

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2017 年 7 月 13 日 (13.07.2017)



(10) 国际公布号
WO 2017/118335 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/112855
- (22) 国际申请日: 2016 年 12 月 29 日 (29.12.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610009341.4 2016 年 1 月 7 日 (07.01.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 陈绪 (CHEN, Xu); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。余晋 (YU, Jin); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。李小龙 (LI, Xiaolong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。丁轶 (DING, Yi); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团

法务部, Zhejiang 311121 (CN)。熊怀东 (XIONG, Huaidong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE,

[见续页]

(54) Title: MAPPING METHOD AND DEVICE

(54) 发明名称: 一种映射方法和设备

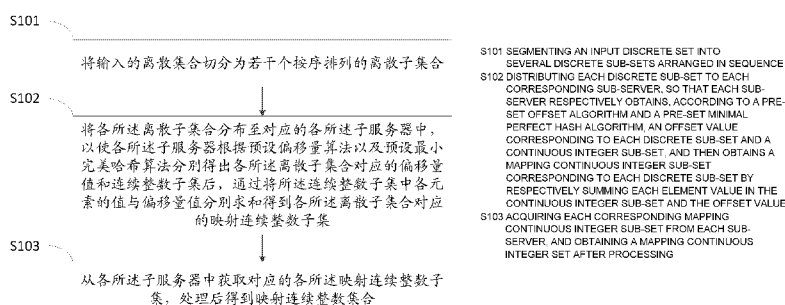


图 1

(57) Abstract: A mapping method and device, which are applied to a primary server in a trunked system, wherein the trunked system further comprises various sub-servers. The method comprises: segmenting a received discrete set into several discrete sub-sets (S101); distributing each discrete sub-set to each corresponding sub-server, so that each sub-server respectively obtains, according to a pre-set offset calculation formula and a minimal perfect hash algorithm, an offset value corresponding to each discrete sub-set and a continuous integer sub-set, and then obtains a mapping continuous integer sub-set corresponding to each discrete sub-set by respectively summing each element value in the continuous integer sub-set and the offset value (S102); and acquiring each corresponding mapping continuous integer sub-set from each sub-server, and obtaining a mapping continuous integer set after combination (S103). The present method is not limited by a stand-alone memory and computing resources, thereby saving hardware resources, and can be used to perform corresponding linear expansion on an input discrete set, thereby improving the mapping conversion efficiency and the learning effect of a machine learning algorithm.

(57) 摘要:

[见续页]



WO 2017/118335 A1



IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,
RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD,
TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种映射方法和设备，应用于集群系统中的主服务器，集群系统还包括各子服务器，包括：将所接收的离散集合切分为若干个离散子集合（S101）；将各离散子集合分布至对应的各子服务器，以使各子服务器根据预设偏移量计算公式以及最小完美哈希算法分别得出各离散子集合对应的偏移量值和连续整数子集后，通过将连续整数子集中各元素与偏移量值分别求和得到各离散子集合对应的映射连续整数子集（S102）；从各子服务器中获取对应的各映射连续整数子集，合并后得到连续整数映射集合（S103）。本方法不会受到单机内存和计算资源的限制，节省了硬件资源，可以对输入地离散集合做相应地线性扩展，提升了映射的转换效率以及机器学习算法的学习效果。

一种映射方法和设备

本申请要求 2016 年 01 月 07 日递交的申请号为 201610009341.4、发明名称为“一种映射方法和设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本发明涉及通信技术领域，特别涉及一种映射方法和设备。

背景技术

随着网络技术的不断发展，互联网领域产生的数据量发生了爆炸式的增长，超大规模模的互联网数据中凌乱地分布着诸多极具意义的信息，通常利用机器学习算法对行业所需的数据信息进行处理与挖掘。尤其是在基于搜索查询结果排序、互联网广告点击率预测、商品个性化推荐、语音识别和智能问答等涉及大规模数据处理的系统中，超大规模机器学习算法已成为最重要的技术支撑之一。

机器学习算法通常是对连续的数值矩阵和向量进行运算，这也就要求了输入数据必须是连续数值空间。然而互联网领域的大规模数据一般都是由用户的点击日志、搜索查询日志或者商品购买日志汇总而来，也就是说，绝大部分的互联网数据都是以离散集合的形式存在，比如：

一组用户 ID 的集合：{user_1, user_2, ..., user_n}；

一组商品 ID 的集合：{item_1, item_2, ..., item_n}；

20 一组搜索查询的集合：{“男装”，“高跟鞋”，...}。

因此在执行机器学习算法之前，先要通过连续数值化方法将离散集合转换为机器学习算法可以使用的连续数值空间，即需要有一种可以从离散集合到连续整数集合的映射：

$$f:S \rightarrow N$$

其中，S 是原始的离散集合，N 是映射之后的自然数集合，范围为 $[0, n-1]$, $n=|S|$ 。

25 通过应用上述的映射关系，就可以将原始的离散集合映射为连续整数集合，即完成了从样本矩阵到数值矩阵的转换，然后将数值矩阵输入到机器学习算法中完成后续的计算过程。

在现有技术的连续数值化方法中，一般都是采用哈希表映射的方式。具体的，首先创建一个哈希表，然后通过查询哈希表判断输入集合中的每一个元素是否已经在哈希表

中在了对应表项。接着根据判断结果选择不同的执行方式，如果对应元素已经在哈希表中在了对应表项，则忽略该元素；如果不存在的话，则为该元素分配一个整数值，这个整数值等于当前哈希表的总元素个数，并同时将这个元素和对应的分配整数值加入当前哈希表中。最终形成的哈希表就是一个映射关系，根据这个映射关系，就可以将原

5 输入集合转换为整数值集合。

在实现本申请的过程中，发明人发现现有技术至少存在如下问题：

(1) 只有将整个原始离散集合的元素存入同一个哈希表中，才能获得全局唯一的整数值，然而单一的哈希表所能够存储的数据容量会受到硬件条件的限制，同时也无法进行并发读写，因此会出现硬件性能无法满足处理要求的问题；

10 (2) 待处理数据无法通过集群资源利用多个进程并行处理，造成处理效率较低，不适合处理现今互联网规模的大规模数据集合；

(3) 哈希表中需要保存原始离散集合的内容作为映射键值，那么如果原始离散集合占据了较大的内存空间，映射键值也会相应地占据较大的内存空间，同时还需要在单机上加载全部的映射对，这些都会使得系统处理原始离散集合规模的上限受到单机内存的

15 上限的限制，而无法进行线性扩展。

以上现有技术所存在的缺点均会不同程度地限制机器学习所需的数据和特征规模，从而会影响机器学习算法最终能取得的效果。

因此，现有技术在针对超大规模离散集合的连续数值化过程时，会受到单机内存和计算资源的限制，对输入集合无法做相应地线性扩展，进而会影响映射的转换效率以及

20 机器学习算法的学习效果，同时也浪费了大量的硬件资源。

发明内容

有鉴于背景技术中的问题，本发明提供一种映射方法，通过对映射算法的优化以及

25 将离散集合切分并行处理的方式，解决现有技术中受到单机内存和计算资源限制的问题，可以对输入的离散集合做相应地线性扩展，节省了硬件资源，同时提升了映射的转换效率以及机器学习算法的学习效果。

该方法应用于集群系统中的主服务器，所述集群系统还包括各子服务器，所述方法包括：

30 将所接收的离散集合切分为若干个按序排列的离散子集合；

将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分别求和得到各所述离散子集合对应的映射连续整数子集；

- 5 从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

优选地，将输入的离散集合切分为若干个离散子集合，具体为：

根据预设哈希函数映射出所述离散集合中各元素的哈希值；

将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

- 10 将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

优选地，处理后得到映射连续整数集合，具体为：

计算出所有各所述映射连续整数子集的并集；

将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

- 15 本发明还提供了一种应用于集群系统中的各子服务器的映射方法，所述集群系统还包括主服务器，该方法包括：

从所述主服务器接收对应的离散子集合；

根据预设偏移量算法以及最小完美哈希算法分别得出所述离散子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集；

- 20

将所述映射连续整数子集转发至所述主服务器，以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

优选地，根据预设偏移量算法得出所述离散子集合对应的偏移量值，具体为：

- 25 判断该离散子集合在所有离散子集合中的所处顺序是否为首位；

若是，则该离散子集合对应的偏移量值为 0；

若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

优选地，根据最小完美哈希算法得出所述离散子集合对应的连续整数子集，具体为：

- 30 根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述

哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列；

根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值；

将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

- 5 优选地，根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，具体为：

通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

- 10 遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

优选地，基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号，具体为：

- 15 根据所述数组以及预设编号计算公式计算出元素对应的编号值；

判断所述编号值是否已被占用；

若否，则所述编号值为所述元素对应的所述哈希函数的编号。

优选地，将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集，具体为：

- 20 根据所述哈希值对应的所述哈希函数的编号，确定在分配该编号之前分配出去的所有编号的个数，所述哈希值对应的整数为所述个数的大小；

将各所述哈希值对应的整数汇总后，得出所述离散子集合对应的连续整数子集。

相应地，本发明提供了一种服务器，所述服务器为应用于处理离散集群系统中的主

- 25 服务器，所述集群系统还包括各子服务器，所述服务器包括：

切分模块，将所接收的离散集合切分为若干个按序排列的离散子集合；

分布模块，将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分别求和得到各所述离散子集合对应的映射连续整数子集；

30

第一处理模块，从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

优选地，所述切分模块具体用于：

根据预设哈希函数映射出所述离散集合中各元素的哈希值；

5 将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

优选地，所述第一处理模块具体用于：

计算出所有各所述映射连续整数子集的并集；

将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

10

相应地，本发明还提供了一种服务器，所述服务器为应用于集群系统中的子服务器，所述集群系统还包括主服务器，所述服务器包括：

接收模块，从所述主服务器接收对应的离散子集合；

15 第二处理模块，根据预设偏移量算法以及最小完美哈希算法分别得出所述离散子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集；

转发模块，将所述映射连续整数子集转发至所述主服务器，以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

20 优选地，所述第二处理模块具体用于：

判断该离散子集合在所有离散子集合中的所处顺序是否为首位；

若是，则该离散子集合对应的偏移量值为 0；

若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

25 优选地，所述第二处理模块还用于：

根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列；

根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值；

30 将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

优选地，所述第二处理模块还用于：

通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

5 遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

优选地，所述第二处理模块还用于：

根据所述数组以及预设编号计算公式计算出元素对应的编号值；

10 判断所述编号值是否已被占用；

若否，则所述编号值为所述元素对应的所述哈希函数的编号。

优选地，其特征在于，所述第二处理模块还用于：

根据所述哈希值对应的所述哈希函数的编号，确定在分配该编号之前分配出去的所有编号的个数，所述哈希值对应的整数为所述个数的大小；

15 将各所述哈希值对应的整数汇总后，得出所述离散子集合对应的连续整数子集。

由此可见，通过应用本申请的技术方案，现有技术针对超大规模离散集合的连续数值化过程时，通过对离散集合切分后利用集群系统中的多台服务器进行并行处理，且设计了最小完美哈希算法和偏移量的映射算法优化方式。以此可以对输入的离散集合做
20 相应地线性扩展，同时提升了映射的转换效率以及机器学习算法的学习效果，并节省了大量的硬件资源。

附图说明

图 1 为本申请提出的一种映射方法的流程示意图；

25 图 2 为本申请提出的一种映射方法的流程示意图；

图 3 为本申请的具体实施例所提出的一种映射方法的流程示意图；

图 4 为本申请提出的一种服务器的结构示意图；

图 5 为本申请提出的一种服务器的结构示意图。

30 具体实施方式

有鉴于背景技术中的问题，本发明提供一种映射方法，通过对映射算法的优化以及将离散集合切分并行处理的方式，解决现有技术中受到单机内存和计算资源限制的问题，可以对输入的离散集合做相应地线性扩展，节省了硬件资源，同时提升了映射的转换效率以及机器学习算法的学习效果。

5 该方法应用于集群系统中的主服务器，所述集群系统还包括各子服务器。

如图 1 所示，为本申请提出的映射方法的流程示意图，包括以下步骤：

S101：将输入的离散集合切分为若干个按序排列的离散子集合。

在本申请的实施方式中，采用如下步骤进行切分：

a) 根据预设哈希函数映射出所述离散集合中各元素的哈希值；

10 b) 将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

c) 将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

其中，在本申请的具体实施方式中，预设正整数一般选取一个较大的质数。

需要说明的是，本申请需要得到的是离散集合拆分后的离散子集合，本申请的保护范围并不限于集合切分的方法，也就是说，进行以上集合的切分方法仅为本申请优选实施例提出的示例，在此基础上还可以选择其他方式来进行切分，以使本申请适用于更多的应用领域，这些改进都属于本发明的保护范围。

15

S102：将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分别求和得到各所述离散子集合对应的映射连续整数子集。

20

在本申请的实施方式中，采用多个子服务器来分配多个离散子集合，对各个离散子集合进行并行处理。

25 S103：从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

在本申请的实施方式中，在获取全部子服务器输出的所有映射连续整数子集后，继续采用如下步骤进行处理：

a) 计算出所有各所述映射连续整数子集的并集；

30 b) 将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

本发明还提供了一种应用于集群系统中的各子服务器的映射方法，所述集群系统还包括主服务器。

如图 2 所示，为本申请提出的映射方法的流程示意图，包括以下步骤：

5 S201：从所述主服务器接收对应的离散子集合。

在本申请的实施方式中，在主服务器将输入的离散集合切分后，各子服务器分别接收各自对应的离散子集合，从而实现了对各离散子集合进行并行处理这一目的。

10 S202：根据预设偏移量算法以及最小完美哈希算法分别得出所述离散子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集。

在本申请的实施方式中，每个连续整数子集中的元素需要分别与对应的偏移量进行求和。举例来说，离散子集合 1、离散子集合 2 和离散子集合 3 分别对应于连续整数子集 1 为{1,2,3,4}，连续整数子集 2 为{1,2,3,4}，连续整数子集 3 为{1,2,3,4}，如果主服务器直接将连续整数子集 1、连续整数子集 2 和连续整数子集 3 进行合并后的映射连续整数集合为{1,2,3,4, 1,2,3,4, 1,2,3,4}，很明显无法实现。故本申请引出了偏移量这一概念，离散子集和 1 的偏移量为 0，离散子集和 2 的偏移量为 4，离散子集和 3 的偏移量为 8，每个连续整数子集中的元素分别通过与对应的偏移量进行求和后得到的对应的映射连续整数子集，则映射连续整数子集 1 为{1,2,3,4}，映射连续整数子集 2 为{5,6,7,8}，映射连续整数子集 3 为{9,10,11,12}，如果主服务器将映射连续整数子集 1、映射连续整数子集 2 和映射连续整数子集 3 进行合并后的映射连续整数集合为{1,2,3,4,5,6,7,8,9,10,11,12}，从而实现了映射结果为连续整数集合这一技术效果。

故本申请具体实施方式中公开了以下偏移量值的计算步骤：

- a) 判断该离散子集合在所有离散子集合中的所处顺序是否为首位；
- 25 b) 若是，则该离散子集合对应的偏移量值为 0；
- c) 若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

需要说明的是，本申请需要实现各映射连续整数子集在合并后仍然为连续整数的集合，故提出一种偏移量的计算方法，本申请的保护范围并不限于上述计算方法，也就是说，进行以上偏移量的计算方法仅为本申请优选实施例提出的示例，在此基础上还可以

30

选择其他方式来进行计算，以使本申请适用于更多的应用领域，这些改进都属于本发明的保护范围。

另外，最小完美哈希算法可以得到离散子集合的连续整数子集，离散子集合中的元素个数和连续整数子集的元素个数是相等的，同时是一一对应且不冲突的。举例来说，如果离散子集合中包含 5 个离散元素，通过最小完美哈希算法后则会形成类似{0,1,2,3,4}这种包含 5 个连续整数的连续整数子集，再通过与对应偏移量之间进行求和后得到离散子集合所对应的映射连续整数子集。

在本申请具体实施方式中，公开了以下最小完美哈希算法的计算步骤：

a) 根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列。

具体的，通过举例来进行说明，如果离散子集合 S_i 中有 4 个元素，分别为 x_1 、 x_2 、 x_3 与 x_4 ，则构造出 4 个哈希函数{ h_0 , h_1 , h_2 , h_4 }。

b) 根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值。

其中，其中编号是通过如下步骤进行确定的：

1)通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

2)分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

3)遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

4)基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

具体的，基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号，包括如下步骤：

①根据所述数组以及预设编号计算公式计算出元素对应的编号值；

②判断所述编号值是否已被占用；

③若否，则所述编号值为所述元素对应的所述哈希函数的编号。

c) 将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

具体的，将各所述哈希值进行排序，包括如下步骤：

a) 根据所述哈希值对应的所述哈希函数的编号, 确定在分配该编号之前分配出去的所有编号的个数, 所述哈希值对应的整数为所述个数的大小;

b) 将各所述哈希值对应的整数汇总后, 得出所述离散子集合对应的连续整数子集。

需要说明的是, 本申请进行以上基于最小完美哈希算法得到离散子集合的连续整数子集的计算过程仅为本申请优选实施例提出的示例, 在此基础上还可以选择其他方式来进行计算, 以使本申请适用于更多的应用领域, 这些改进都属于本发明的保护范围。

S203: 将所述映射连续整数子集转发至所述主服务器, 以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

由以上内容可知, 通过应用本申请的技术方案, 在针对现有技术针对超大规模离散集合的连续数值化过程时, 通过对离散集合切分后利用集群系统中的多台服务器进行并行处理, 且设计了最小完美哈希算法和偏移量的映射算法优化方式。以此可以对输入的离散集合做相应地线性扩展, 并使得在生成的映射关系中不需要保存原始离散集合的信息, 显著降低了内存占用, 同时提升了映射的转换效率以及机器学习算法的学习效果, 并节省了大量的硬件资源。

为了进一步阐述本发明的技术思想, 现结合图 3 所示的具体的应用场景, 对本发明的技术方案进行说明。

在此具体的应用场景中, 提出了一种映射方法。该方法包括如下步骤:

步骤 1: 接收输入的离散集合, 预先选定一个哈希函数 h , 通过该哈希函数映射出所述离散集合中各元素的哈希值, 将各所述哈希值对正整数 k 取模得到各所述元素的哈希值所对应的模值, 将模值相等的元素分入同一个离散子集合, 以切分成 k 个离散子集合。

在本实施方式中, 步骤 1 中的第 i 个离散子集合 S_i ($1 \leq i \leq k$) 可以表示为:

$$S_i = \{x, h(x) \bmod k = i\}$$

其中, x 为离散子集合中的元素, $h(x)$ 为元素 x 对应的哈希值, i 的范围为 $[1, k]$ 。

通过步骤 1 切分得到的各个离散子集合中没有重复元素, 且各个离散子集合的规模也基本均等, 然后通过将各离散子集合分布至集群系统中对应的各子服务器中, 每台子服务器可以并行处理各自对应的离散子集合。

也就是说，步骤 1 是将离散集合中所有基于哈希值取模后模值为 i 的元素分入离散子集合 S_i 中。

步骤 2：各子服务器对各自对应的离散子集合并行处理，计算出各离散子集合的偏移量值，偏移量(Offset)的递推定义如下：

$$\begin{cases} Offset_1 = 0 \\ Offset_i = \sum_{j=1}^{i-1} |S_j|, 1 < i \leq k \end{cases}$$

在本实施方式中， $Offset_i$ 为第 i 个离散子集合所对应的偏移量值， $|S_j|$ ($1 \leq j \leq i-1$) 为第 j 个离散子集合中的元素个数。

具体的，第一个离散子集合的 $Offset_1$ 偏移量值为 0，从第二个离散子集开始，各离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

步骤 3：各子服务器对各自对应的离散子集合并行处理，对于每个离散子集合子集 S_i ，均基于最小完美哈希算法 (Minimal Perfect Hash) 生成一个映射关系 f_i ：

$$f_i : S_i \rightarrow N_i, |S_i| = n_i, N_i = \{0, 1, \dots, n_i - 1\}$$

其中，这个映射关系 f_i 将离散子集合 S_i 映射到一个连续整数空间集合 N_i 中， N_i 的范围是 $[0, n_i - 1]$ ， $|S_i| = n_i$ 表示第 i 个离散子集合中的元素个数为 n_i 个。

在本实施方式中，步骤 3 中最小完美哈希映射关系的计算步骤如下：

a) 映射步：根据离散子集合 S_i 中元素的个数 n_i ，从一组哈希函数 H 中随机选取并构造 n_i 个哈希函数 $\{h_0, h_1, \dots, h_{n_i-1}\}$ ，构造的哈希函数的个数是和离散子集合中的元素个数是相等的。选取已知哈希函数 h' ，分别为离散子集合 S_i 中的任意元素 x 生成 n_i 个哈希值 $h_0', h_1', \dots, h_{n_i-1}'$ ，从而有：

$$\begin{aligned} h_0 &= h_0' \bmod \eta \\ h_1 &= h_1' \bmod \eta + \eta \\ h_2 &= h_2' \bmod \eta + 2\eta \\ &\dots \end{aligned}$$

以此类推，就得到了关于元素 x 的 n_i 个哈希函数，离散子集合中的所有元素均通过上述规则进行处理。其中， η 是一个预先设置的参数，通过上述方式选取出来的哈希函数的值域是 $[0, \eta \times n_i)$ ，也就是说，针对离散子集合 S_i 中的 n_i 个元素，这组哈希函数 $\{h_0,$

$h_1, \dots, h_{n_i-1}$ 的输出值个数为 $n_i \times n_i$ 个。

创建一个无环 n_i 部超图 (acyclic n_i -partite hypergraph), 超图的每个独立子集边数和 S_i 的元素个数 n_i 相同, 超图的每个节点对应一个由上面生成的 n_i 个哈希函数对子集合中元素计算得到的输出值, 输出值的范围是 $[0, m-1]$, 这样的节点有 m 个, 其中 $m = n_i \cdot n_i$ 。

5 b) 分配步: 在前面创建的无环 n_i 部超图中, 离散子集合 S_i 中的任意元素 x 是由 n_i 个哈希函数输出值对应到 n_i 个节点, 可以表示为 $V = \{v_0, v_1, \dots, v_{n_i-1}\}$, 每个节点上对应有一个整数值, 离散子集合 S_i 中的任意元素 x 分配整数值的步骤如下:

1) 遍历所述无环超图的每一条边, 在这条边上, 找到在每条上第一个还没有被分配的节点 u , 令

$$10 \quad g[u] = (j - \sum_{v \in e \wedge \text{Visited}[v] = \text{ture}} g[v]) \bmod 3$$

根据上述节点计算公式得出各所述节点对应的计算结果, 以形成基于计算结果的数组 $g = \{g_0, g_1, \dots, g_{m-1}\}$, 其中 $0 \leq g_i \leq n_i$ 。数组 $g = \{g_0, g_1, \dots, g_{m-1}\}$ 适用于离散子集合 S_i 中的任意元素 x 的计算过程。

2) 根据所述数组 $g = \{g_0, g_1, \dots, g_{m-1}\}$ 以及预设编号计算公式计算出元素对应的编号值, 进而确定离散子集合 S_i 中的任意元素 x 对应到唯一的一个节点上所属的整数值。其中, 编号计算公式如下:

$$i = (g_{h_0(x)} + g_{h_1(x)} + \dots + g_{h_{(n_i-1)}(x)}) \bmod n_i$$

然后判断计算所得的编号值 i 是否已被使用, 若否, 则该编号值为元素 x 对应的所述哈希函数的编号, 即哈希函数 h_i 的对应计算结果为元素 x 对应的整数值, 该整数值的取值范围是 $[0, m)$; 若是, 则顺延找到下一个编号 $i+1$, 判断编号值 $i+1$ 是否已被使用, 若否, 则编号值 $i+1$ 为所述元素对应的所述哈希函数的编号, 即哈希函数 h_{i+1} 的对应计算结果为元素 x 对应的整数值, 该整数值的取值范围是 $[0, m)$, 依此类推。

20 c) 排序步: 分配步已经为离散子集合中的每个元素分配了一个整数值, 整数值的取值范围是 $[0, m)$, 为了得到最小完美哈希函数, 需要将整数值的取值范围是 $[0, m)$ 缩小至 $[0, n_i-1]$ 。具体步骤如下:

生成一个序号表, 其中序号表是一个长度为 n_i 的一维数组, 其中每个下标对应的值表示, 在这个下标之前被之前分配步使用过的整数个数。具体可参考如下的排序公式:

$$\begin{cases} rank[0]=0 \\ rank[i]=rank[i-1]+assigned[i-1], 1 \leq i < n_i \end{cases}$$

其中, $assigned[i]$ 表示第 i 个数是否在分配步被使用。经过排序步, 离散子集中的元素就被一一映射到连续的整数空间子集合中, 该整数空间集合取值范围是 $[0, n_i-1]$ 。最小完美哈希函数可以通过如下公式进行表示:

$$5 \quad mph_i(x) = rank[h_i(x)]$$

其中, $mph_i(x)$ 为第 i 个离散子集合 S_i 中任意元素 x 对应的最小完美哈希函数的输出值, $rank[h_i(x)]$ 为排序步的具体处理过程。

步骤 4: 各子服务器并行处理, 基于步骤 3 得出的连续的整数空间子集合, 将每个子服务器的整数空间集合中各元素的哈希值分别加上步骤 2 计算出的对应的偏移量值, 得到最终的映射连续整数子集合。

在本实施方式中, 最终的映射连续整数子集合可以表示为:

$$f_i(x) = mph_i(x) + Offset_i$$

其中, $mph_i(x)$ 为第 i 个离散子集合 S_i 中任意元素 x 对应的最小完美哈希函数的输出值, $Offset_i$ 为第 i 个离散子集合所对应的偏移量值。

步骤 5: 将各子服务器中生成的映射连续整数子集合汇总到一个集合, 形成最终输出的映射连续整数集合。

需要说明的是, 本具体实施方式所涉及的映射方法仅为本申请优选实施例提出的示例, 在此基础上还可以选择其他类似的方式来进行计算以得到类似的结果, 以使本申请适用于更多的应用领域, 这些改进都属于本发明的保护范围。

由以上内容可知, 本具体实施方式在针对现有技术针对超大规模离散集合的连续数值化过程时, 通过对离散集合切分后利用集群系统中的多台服务器进行并行处理, 且设计了最小完美哈希算法和偏移量的映射算法优化方式。以此可以对输入的离散集合做相应地线性扩展, 并使得在生成的映射关系中不需要保存原始离散集合的信息, 显著降低了内存占用, 同时提升了映射的转换效率以及机器学习算法的学习效果, 并节省了大

量的硬件资源。

为达到以上技术目的，相应地，本申请还提出了一种服务器，所述服务器为应用于处理离散集群系统中的主服务器，所述集群系统还包括各子服务器，如图 4 所示，所述

5 服务器包括：

切分模块 401，将所接收的离散集合切分为若干个按序排列的离散子集合；

分布模块 402，将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分
10 别求和得到各所述离散子集合对应的映射连续整数子集；

第一处理模块 403，从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

在具体的应用场景中，所述切分模块具体用于：

根据预设哈希函数映射出所述离散集合中各元素的哈希值；

15 将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

在具体的应用场景中，所述第一处理模块具体用于：

计算出所有各所述映射连续整数子集的并集；

将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

20

为达到以上技术目的，相应地，本申请还提出了一种服务器，所述服务器为应用于集群系统中的子服务器，所述集群系统还包括主服务器，如图 5 所示，所述服务器包括：

接收模块 501，从所述主服务器接收对应的离散子集合；

第二处理模块 502，根据预设偏移量算法以及最小完美哈希算法分别得出所述离散
25 子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集；

转发模块 503，将所述映射连续整数子集转发至所述主服务器，以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

30 在具体的应用场景中，所述第二处理模块具体用于：

判断该离散子集合在所有离散子集合中的所处顺序是否为首位；

若是，则该离散子集合对应的偏移量值为 0；

若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

5 在具体的应用场景中，所述第二处理模块还用于：

根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列；

根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值；

10 将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

在具体的应用场景中，所述第二处理模块还用于：

通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

15 遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

在具体的应用场景中，所述第二处理模块还用于：

根据所述数组以及预设编号计算公式计算出元素对应的编号值；

20 判断所述编号值是否已被占用；

若否，则所述编号值为所述元素对应的所述哈希函数的编号。

在具体的应用场景中，所述第二处理模块还用于：

根据所述哈希值对应的所述哈希函数的编号，确定在分配该编号之前分配出去的所有编号的个数，所述哈希值对应的整数为所述个数的大小；

25 将各所述哈希值对应的整数汇总后，得出所述离散子集合对应的连续整数子集。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到本发明可以通过硬件实现，也可以借助软件加必要的通用硬件平台的方式来实现。基于这样的理解，本发明的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质（可以是 CD-ROM，U 盘，移动硬盘等）中，包括若干指令用以使得一台计算

30

机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施场景所述的方法。

本领域技术人员可以理解附图只是一个优选实施场景的示意图，附图中的模块或流程并不一定是实施本发明所必须的。

- 5 本领域技术人员可以理解实施场景中的装置中的模块可以按照实施场景描述进行分布于实施场景的装置中，也可以进行相应变化位于不同于本实施场景的一个或多个装置中。上述实施场景的模块可以合并为一个模块，也可以进一步拆分成多个子模块。

上述本发明序号仅仅为了描述，不代表实施场景的优劣。

- 10 以上公开的仅为本发明的几个具体实施场景，但是，本发明并非局限于此，任何本领域的技术人员能思之的变化都应落入本发明的保护范围。

权利要求书

1、一种映射方法，其特征在于，所述方法应用于集群系统中的主服务器，所述集群系统还包括各子服务器，所述方法包括：

将输入的离散集合切分为若干个按序排列的离散子集合；

5 将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分别求和得到各所述离散子集合对应的映射连续整数子集；

10 从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

2、如权利要求 1 所述的方法，其特征在于，将所接收的离散集合切分为若干个离散子集合，具体为：

根据预设哈希函数映射出所述离散集合中各元素的哈希值；

将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

15 将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

3、如权利要求 1 所述的方法，其特征在于，处理后得到映射连续整数集合，具体为：

计算出所有各所述映射连续整数子集的并集；

将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

20

4、一种映射方法，其特征在于，所述方法应用于集群系统中的各子服务器，所述集群系统还包括主服务器，所述方法包括：

从所述主服务器接收对应的离散子集合；

25 根据预设偏移量算法以及最小完美哈希算法分别得出所述离散子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集；

将所述映射连续整数子集转发至所述主服务器，以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

30 5、如权利要求 4 所述的方法，其特征在于，根据预设偏移量算法得出所述离散子

集合对应的偏移量值，具体为：

判断该离散子集合在所有离散子集合中的所处顺序是否为首位；

若是，则该离散子集合对应的偏移量值为 0；

若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的
5 元素个数的总和。

6、如权利要求 4 所述的方法，其特征在于，根据最小完美哈希算法得出所述离散子集合对应的连续整数子集，具体为：

根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列；

10 根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值；

将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

7、如权利要求 6 所述的方法，其特征在于，根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，具体为：

15 通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

20 基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

8、如权利要求 7 所述的方法，其特征在于，基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号，具体为：

根据所述数组以及预设编号计算公式计算出元素对应的编号值；

判断所述编号值是否已被占用；

25 若否，则所述编号值为所述元素对应的所述哈希函数的编号。

9、如权利要求 6 或 8 所述的方法，其特征在于，将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集，具体为：

根据所述哈希值对应的所述哈希函数的编号，确定在分配该编号之前分配出去的所有编号的个数，所述哈希值对应的整数为所述个数的大小；

30 将各所述哈希值对应的整数汇总后，得出所述离散子集合对应的连续整数子集。

10、一种服务器，其特征在于，所述服务器为应用于集群系统中的主服务器，所述集群系统还包括各子服务器，所述服务器包括：

切分模块，将输入的离散集合切分为若干个按序排列的离散子集合；

5 分布模块，将各所述离散子集合分布至对应的各所述子服务器中，以使各所述子服务器根据预设偏移量算法以及预设最小完美哈希算法分别得出各所述离散子集合对应的偏移量值和连续整数子集后，通过将所述连续整数子集中各元素的值与偏移量值分别求和得到各所述离散子集合对应的映射连续整数子集；

10 第一处理模块，从各所述子服务器中获取对应的各所述映射连续整数子集，处理后得到映射连续整数集合。

11、如权利要求 10 所述的服务器，其特征在于，所述切分模块具体用于：

根据预设哈希函数映射出所述离散集合中各元素的哈希值；

将各所述哈希值对预设正整数取模得到各所述元素的哈希值所对应的模值；

将模值相等的元素分入同一个离散子集合，以形成预设正整数个所述离散子集合。

15 12、如权利要求 10 所述的服务器，其特征在于，所述第一处理模块具体用于：

计算出所有各所述映射连续整数子集的并集；

将并集后集合中所有的元素按照大小顺序排列后得到映射连续整数集合。

20 13、一种服务器，其特征在于，所述服务器为应用于集群系统中的子服务器，所述集群系统还包括主服务器，所述服务器包括：

接收模块，从所述主服务器接收对应的离散子集合；

第二处理模块，根据预设偏移量算法以及最小完美哈希算法分别得出所述离散子集合对应的偏移量值和连续整数子集后，将所述连续整数子集中各元素的值与偏移量值分别求和得到所述离散子集合对应的映射连续整数子集；

25 转发模块，将所述映射连续整数子集转发至所述主服务器，以使所述主服务器将该映射连续整数子集以及所有从其他子服务器中获取的映射连续整数子集进行处理后得到映射连续整数集合。

14、如权利要求 13 所述的服务器，其特征在于，所述第二处理模块具体用于：

判断该离散子集合在所有离散子集合中的所处顺序是否为首位；

30 若是，则该离散子集合对应的偏移量值为 0；

若否，则该离散子集合对应的偏移量值为所处顺序在其之前的所有离散子集合中的元素个数的总和。

15、如权利要求 13 所述的服务器，其特征在于，所述第二处理模块还用于：

根据该离散子集合中元素的个数，构造出对应个数且带有编号的哈希函数，各所述
5 哈希函数的编号形成了一个从 0 开始的连续正整数的数字序列；

根据预设编号分配策略确定各所述元素对应的所述哈希函数的编号，并分别得出各所述元素对应的各所述哈希值；

将各所述哈希值进行排序，以得出所述离散子集合对应的连续整数子集。

16、如权利要求 15 所述的服务器，其特征在于，所述第二处理模块还用于：

10 通过各所述元素基于各所述哈希函数的全部映射结果，确定该离散子集合对应的所有哈希值个数；

分别以所述元素个数和所述哈希值个数为边数和节点数，构造无环超图；

遍历所述无环超图的每一条边，根据预设节点计算公式得出各所述节点对应的计算结果，以形成基于计算结果的数组；

15 基于数组以及预设编号计算公式，确定各所述元素对应的所述哈希函数的编号。

17、如权利要求 16 所述的服务器，其特征在于，所述第二处理模块还用于：

根据所述数组以及预设编号计算公式计算出元素对应的编号值；

判断所述编号值是否已被占用；

若否，则所述编号值为所述元素对应的所述哈希函数的编号。

20 18、如权利要求 13 或 17 所述的服务器，其特征在于，所述第二处理模块还用于：

根据所述哈希值对应的所述哈希函数的编号，确定在分配该编号之前分配出去的所有编号的个数，所述哈希值对应的整数为所述个数的大小；

将各所述哈希值对应的整数汇总后，得出所述离散子集合对应的连续整数子集。

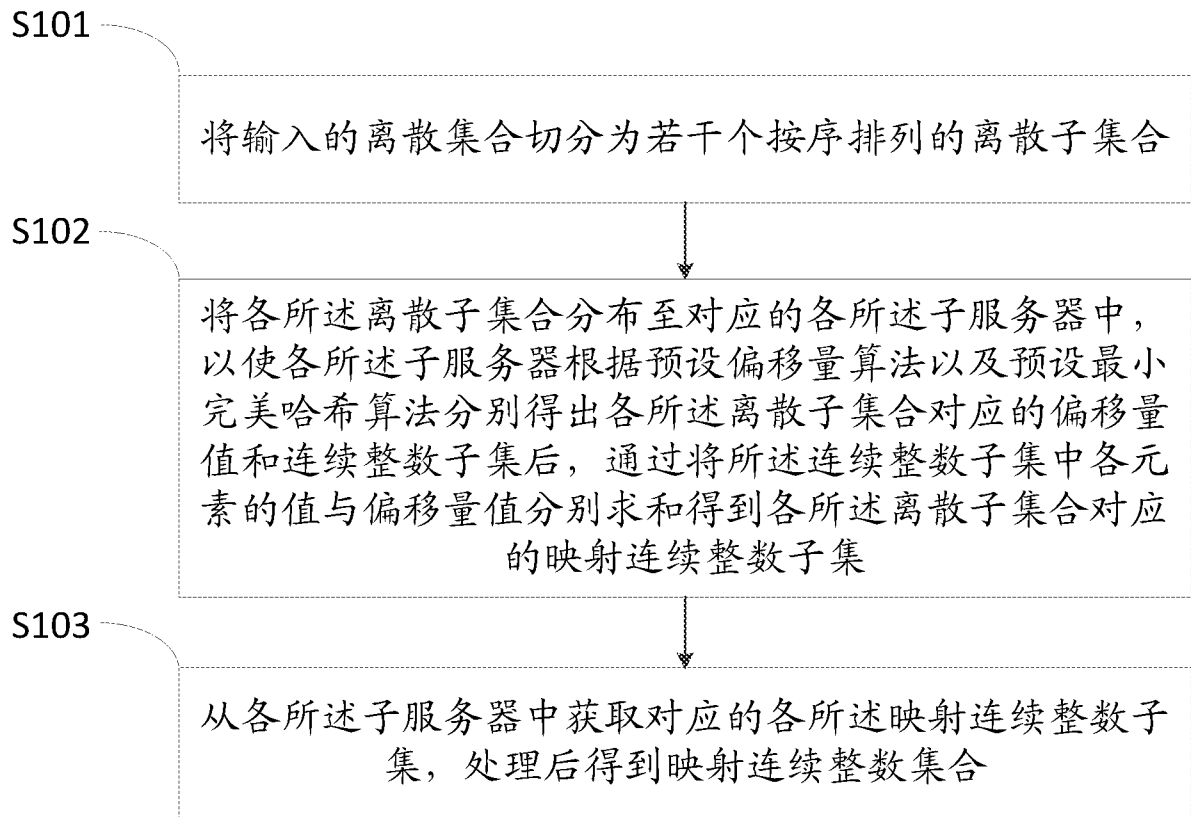


图 1

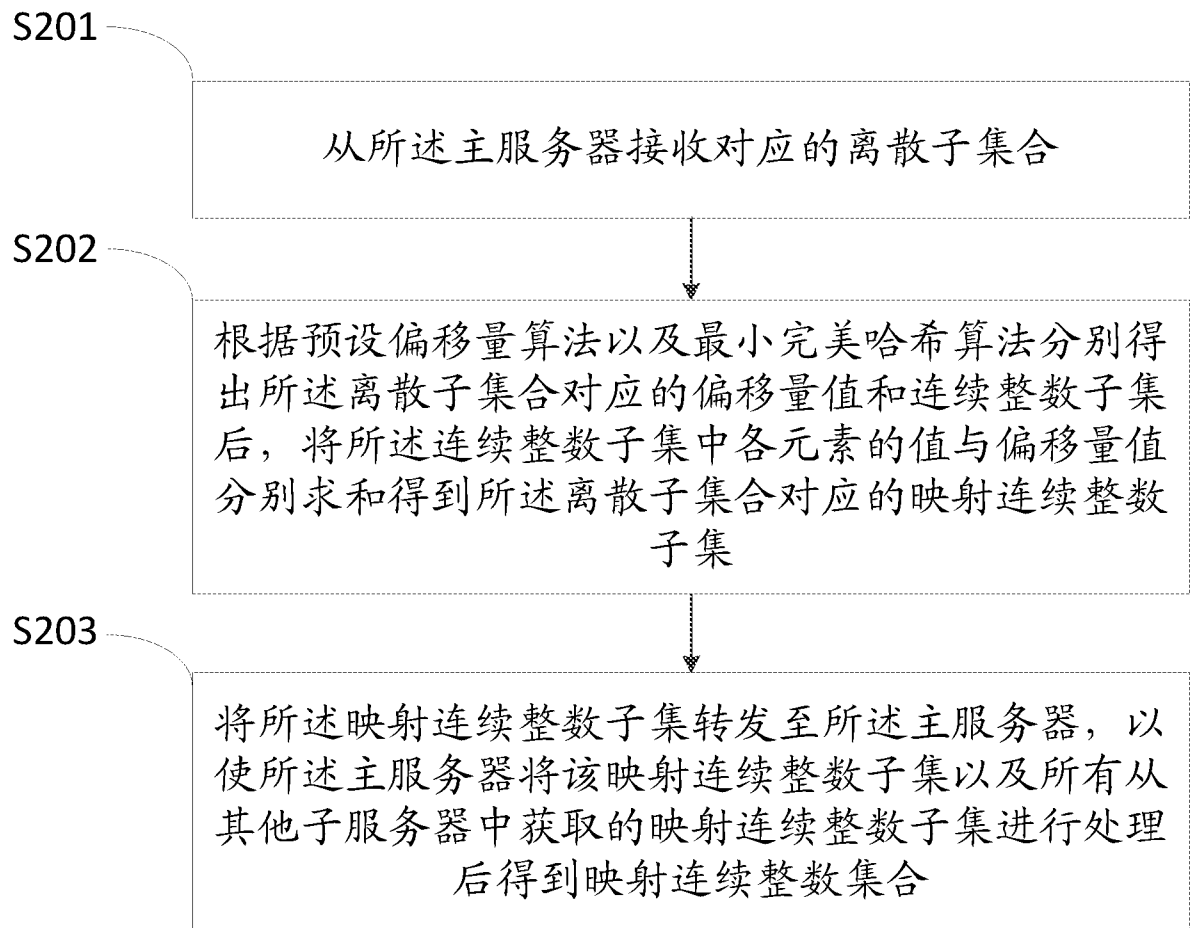


图 2

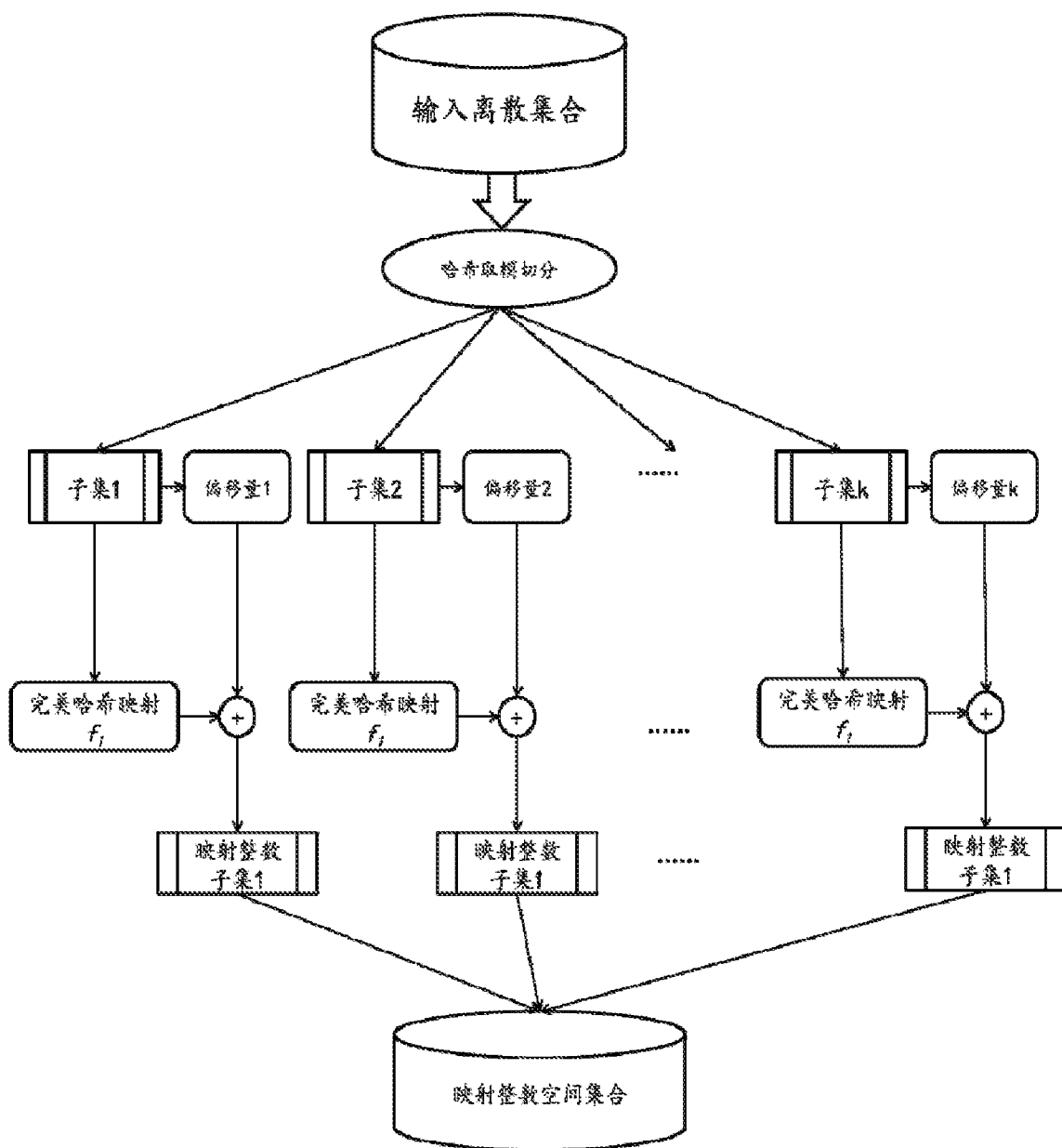


图 3

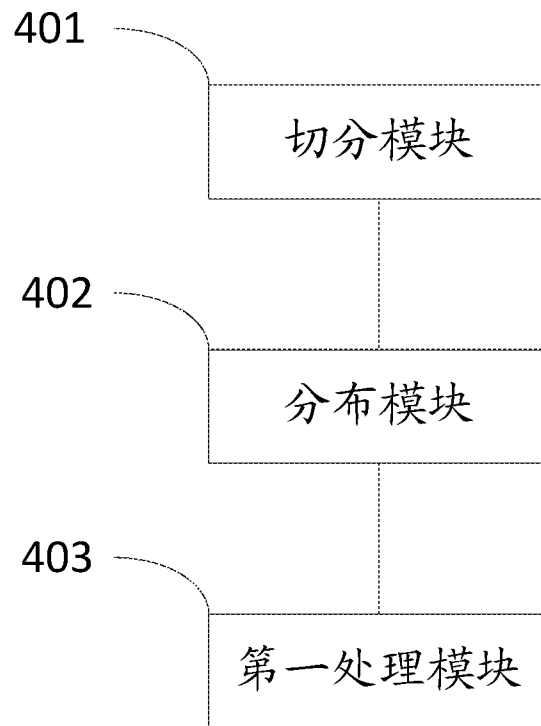


图 4

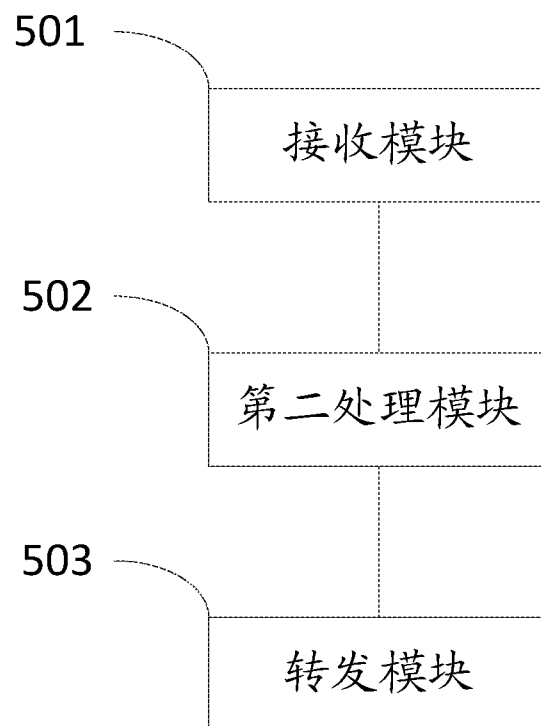


图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/112855

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE, GOOGLE: offset, cluster, server, map, discrete, continuous, set, hash, migration, plus

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102298633 A (XIAMEN MEIYA PICO INFORMATION CO., LTD.) 28 December 2011 (28.12.2011) description, paragraphs [0007]-[0011]	1-18
A	CN 101038590 A (WUHAN UNIVERSITY) 19 September 2007 (19.09.2007) the whole document	1-18
A	US 2010235606 A1 (ORACLE AMERICA, INC.) 16 September 2010 (16.09.2010) the whole document	1-18
A	US 2015293984 A1 (HITACHI DATA SYSTEMS CORPORATION) 15 October 2015 (15.10.2015) the whole document	1-18

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 07 February 2017	Date of mailing of the international search report 07 April 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer KONG, Xin Telephone No. (86-10) 62413668

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/112855

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102298633 A	28 December 2011	None	
CN 101038590 A	19 September 2007	None	
US 2010235606 A1	16 September 2010	EP 2406736 A2	18 January 2012
		CN 102395962 A	28 March 2012
		WO 2010104902 A2	16 September 2010
US 2015293984 A1	15 October 2015	JP 2016505935 A	25 February 2016
		EP 2962218 A1	06 January 2016
		CN 104813321 A	29 July 2015
		WO 2014133497 A1	04 September 2014

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, WPI, EPODOC, CNKI, IEEE, GOOGLE: 集群, 服务器, 映射, 离散, 集合, 哈希, 偏移, 加, 连续, cluster, server, map, discrete, continuous, set, hash, migration, plus																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 102298633 A (厦门市美亚柏科信息股份有限公司) 2011年 12月 28日 (2011 - 12 - 28) 说明书第[0007]-[0011]段</td> <td>1-18</td> </tr> <tr> <td>A</td> <td>CN 101038590 A (武汉大学) 2007年 9月 19日 (2007 - 09 - 19) 全文</td> <td>1-18</td> </tr> <tr> <td>A</td> <td>US 2010235606 A1 (ORACLE AMERICA, INC.) 2010年 9月 16日 (2010 - 09 - 16) 全文</td> <td>1-18</td> </tr> <tr> <td>A</td> <td>US 2015293984 A1 (HITACHI DATA SYSTEMS CORPORATION) 2015年 10月 15日 (2015 - 10 - 15) 全文</td> <td>1-18</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 102298633 A (厦门市美亚柏科信息股份有限公司) 2011年 12月 28日 (2011 - 12 - 28) 说明书第[0007]-[0011]段	1-18	A	CN 101038590 A (武汉大学) 2007年 9月 19日 (2007 - 09 - 19) 全文	1-18	A	US 2010235606 A1 (ORACLE AMERICA, INC.) 2010年 9月 16日 (2010 - 09 - 16) 全文	1-18	A	US 2015293984 A1 (HITACHI DATA SYSTEMS CORPORATION) 2015年 10月 15日 (2015 - 10 - 15) 全文	1-18
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	CN 102298633 A (厦门市美亚柏科信息股份有限公司) 2011年 12月 28日 (2011 - 12 - 28) 说明书第[0007]-[0011]段	1-18															
A	CN 101038590 A (武汉大学) 2007年 9月 19日 (2007 - 09 - 19) 全文	1-18															
A	US 2010235606 A1 (ORACLE AMERICA, INC.) 2010年 9月 16日 (2010 - 09 - 16) 全文	1-18															
A	US 2015293984 A1 (HITACHI DATA SYSTEMS CORPORATION) 2015年 10月 15日 (2015 - 10 - 15) 全文	1-18															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。 <table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件													
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																
国际检索实际完成的日期 2017年 2月 7日		国际检索报告邮寄日期 2017年 4月 7日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		受权官员 孔昕 电话号码 (86-10) 62413668															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/112855

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	102298633	A	2011年 12月 28日	无			
CN	101038590	A	2007年 9月 19日	无			
US	2010235606	A1	2010年 9月 16日	EP	2406736	A2	2012年 1月 18日
				CN	102395962	A	2012年 3月 28日
				WO	2010104902	A2	2010年 9月 16日
US	2015293984	A1	2015年 10月 15日	JP	2016505935	A	2016年 2月 25日
				EP	2962218	A1	2016年 1月 6日
				CN	104813321	A	2015年 7月 29日
				WO	2014133497	A1	2014年 9月 4日