(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104298680 B
(45)授权公告日 2019.01.11

(21)申请号 201310302711.X

(22)申请日 2013.07.18

(65)同一申请的已公布的文献号
申请公布号 CN 104298680 A

(43)申请公布日 2015.01.21

(73)专利权人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市福田区振兴路赛格科技园2栋东403室

(72)发明人 王才平

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 贾允 肖丁

(51)Int.Cl.
G06F 16/901(2019.01)

(56)对比文件
CN 102073712 A,2011.05.25,
CN 102222085 A,2011.10.19,
US 2013144882 A1,2013.06.06,
US 2013103713 A1,2013.04.25,
CN 102043795 A,2011.05.04,

审查员 王博实

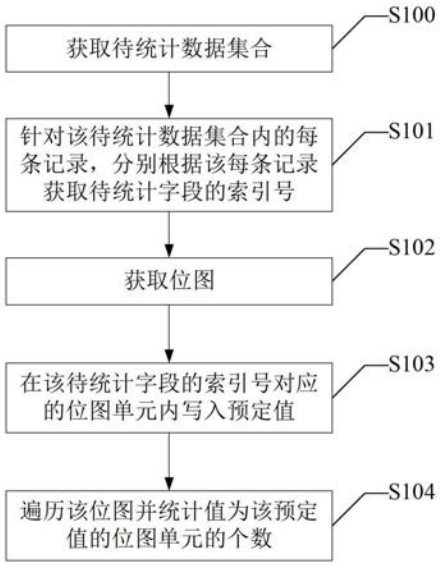
权利要求书2页 说明书6页 附图8页

(54)发明名称

数据统计方法及数据统计装置

(57)摘要

一种数据统计方法，包括：获取待统计数据集合；针对待统计数据集合内的每条记录，分别根据每条记录获取待统计字段的索引号；获取位图，该位图包括多个位图单元，每个位图单元具有唯一的索引号；在与该待统计字段的索引号对应的位图单元内写入预定值；遍历该位图并统计值为预定值的位图单元的个数。此外，本发明还提供一种数据统计装置。上述的数据统计方法及数据统计装置具有更高地统计效率及更低的内存消耗。



CN 104298680 B

1.一种数据统计方法,其特征在于,包括:

获取待统计数据集合;

针对该待统计数据集合内的每条记录,分别根据该每条记录获取待统计字段的索引号;

获取位图,该位图包括多个位图单元,每个位图单元具有唯一的索引号;所述获取位图包括:创建该位图,该位图中位图单元的数量不少于该待统计字段可能值的数量;

在与该待统计字段的索引号对应的位图单元内写入预定值;

遍历该位图并统计值为该预定值的位图单元的个数。

2.如权利要求1所述的方法,其特征在于,根据该每条记录获取待统计字段的索引号包括:

若该待统计字段为无符号整形则将该待统计字段的值作为该待统计字段的索引号,否则将该待统计字段的值按一一对应的方式映射为无符号整形数,并将映射得到的无符号整形数作为该待统计字段的索引号。

3.如权利要求1所述的方法,其特征在于,所述获取位图还包括:

根据该待统计字段的索引号N及预定的分段长度L计算对应的分段索引号k=N/L,并对计算结果取整;

判断分段索引号为k的位图分段是否已被创建,若是,则获取分段索引号为k的位图分段;若否,则在预定的位图空间内创建一个新的位图分段,标记该新的位图的分段索引号为k,并记录其创建顺序索引号n。

4.如权利要求3所述的方法,其特征在于,所述在与该待统计字段的索引号对应的位图单元内写入预定值包括:

根据该待统计字段的索引号N及预定的分段长度L计算偏移值offset=N%L;

根据以下公式获取与该待统计字段的索引号对应的位图单元的地址:S0+L*n+offset,其中S0为该位图空间的起始地址;以及

根据获取的地址写入预定值。

5.一种数据统计装置,其特征在于,包括:

数据获取模块,用于获取待统计数据集合;

索引号获取模块,用于针对该数据获取模块获取的该待统计数据集合内的每条记录,分别根据该每条记录获取待统计字段的的索引号;

位图获取模块,用于创建位图,该位图中位图单元的数量不少于该待统计字段可能值的数量;还用于获取位图,该位图包括多个位图单元,每个位图单元具有唯一的索引号;

写入模块,用于在索引号与该每条记录的对应字段的索引号相同的位图单元内写入预定值;

统计模块,用于遍历该位图并统计值为该预定值的位图单元的个数。

6.如权利要求5所述的装置,其特征在于,该索引号获取模块用于:若该待统计字段为无符号整形则将该待统计字段的值作为该待统计字段的索引号,否则将该待统计字段的值按一一对应的方式映射为无符号整形数,并将映射得到的无符号整形数作为该待统计字段的索引号。

7.如权利要求5所述的装置,其特征在于,该位图获取模块包括:

分段索引号获取单元，用于根据该待统计字段的索引号N及预定的分段长度L计算对应的分段索引号k＝N/L，并对计算结果取整；

分段获取模块，用于判断分段索引号为k的位图分段是否已被创建，若是，则获取分段索引号为k的位图分段；若否，则在预定的位图空间内创建一个新的位图分段，标记该新的位图的分段索引号为k，并记录其创建顺序索引号k。

8.如权利要求7所述的装置，其特征在于，该写入模块包括：

偏移值获取单元，用于根据该待统计字段的索引号N及预定的分段长度L计算偏移值offset＝N%L；

地址获取单元，用于根据以下公式获取与该待统计字段的索引号对应的位图单元的地址：S0+L*k+offset，其中S0为该位图空间的起始地址；以及

写入单元，用于根据获取的地址写入预定值。

数据统计方法及数据统计装置

技术领域

[0001] 本发明涉及数据统计技术领域，尤其涉及一种统计方法及装置。

背景技术

[0002] 在数据分析中，往往需要统计某种值集合的唯一值的个数，现有的唯一值统计技术主要是通过利用Java、C++、Python等高级语言编写程序的方式来统计唯一值的个数，例如：利用上述各类高级语言中的Set对象存储账户号码，然后取得Set的size得到唯一值个数，代码可如下示：

[0003] HashSet<String>set＝new HashSet<String>();

[0004] while(...){

[0005] String userId＝xxx;

[0006] set.add(xxx);

[0007] }

[0008] return set.size();

[0009] 上述方法使用简单，代码量少，容易理解。然而在进行海量数据分析时，假如账户号码唯一值非常多，Set就会消耗大量的内存，经常导致程序内存溢出，同时由于通过哈希值计算、存储与散列冲突解决、比较、动态扩充内部存储数组等，都会花费一定代价，故该方式的效率也较低。

发明内容

[0010] 有鉴于此，本发明提供一种数据统计方法及数据统计装置，其可减少内存消耗，同时提升统计效率。

[0011] 一种数据统计方法，包括：获取待统计数据集合；针对该待统计数据集合内的每条记录，分别根据该每条记录获取待统计字段的索引号；获取位图，该位图包括多个位图单元，每个位图单元具有唯一的索引号；在与该待统计字段的索引号对应的位图单元内写入预定值；遍历该位图并统计值为该预定值的位图单元的个数。

[0012] 一种唯一值数量统计装置，包括：数据获取模块，用于获取待统计数据集合；索引号获取模块，用于针对该数据获取模块获取的该待统计数据集合内的每条记录，分别根据该每条记录获取待统计字段的的索引号；位图获取模块，用于获取位图，该位图包括多个位图单元，每个位图单元具有唯一的索引号；写入模块，用于在索引号与该每条记录的对应字段的索引号相同的位图单元内写入预定值；以及统计模块，用于遍历该位图并统计值为该预定值的位图单元的个数。

[0013] 上述的数据统计方法及装置中，通过在位图内的索引号与待统计数据集合内的每条记录的对应字段的索引号相同的位图单元内写入预定值，来统计待统计数据集合中的唯一值，可有效减少内存开销，避免了内存溢出的风险。而且由于减少了哈希值比较的运算，从而提高了统计效率。

[0014] 为让本发明的上述和其他目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附图式，作详细说明如下。

附图说明

[0015] 图1是第一实施例提供的数据统计方法的流程图。

[0016] 图2是第一实施例提供的数据统计方法中的位图的示意图。

[0017] 图3是第二实施例提供的数据统计方法的部分步骤流程图。

[0018] 图4是第三实施例提供的数据统计方法的部分步骤流程图。

[0019] 图5是第三实施例提供的数据统计方法中的分段位图的示意图。

[0020] 图6是本发明实施例的数据统计方法测试数据集示意图。

[0021] 图7是本发明实施例的数据统计方法内存消耗测试结果示意图。

[0022] 图8为本发明实施例的数据统计方法的时间测试结果示意图。

[0023] 图9为第四实施例提供的数据统计装置的结构框图。

[0024] 图10为第四实施例提供的数据统计装置的位图获取模块的结构框图。

[0025] 图11为第四实施例提供的数据统计装置的写入模块的结构框图。

具体实施方式

[0026] 为更进一步阐述本发明为实现预定发明目的所采取的技术手段及功效，以下结合附图及较佳实施例，对依据本发明的具体实施方式、结构、特征及其功效，详细说明如后。

[0027] 图1是第一实施例提供的数据统计方法的流程图。上述的数据统计方法可应用于计算机中，上述的计算机可以是个人电脑或者是服务器。

[0028] 在步骤S100，获取待统计数据集合。

[0029] 待统计数据集合中可包括多条记录，每条记录包含需要统计唯一值的字段内容，该字段内容可以为多种形式，例如：用户帐号、邮箱地址、IP地址、机器物理地址等。在本实施例一具体实施方式中，待统计数据集合可从存储于云端服务器中的SNS(Social Networking Services，社会性网络服务)操作日志中获取，根据统计目的，云端服务器可提取SNS操作日志中与统计目的相关的字段，整合为待统计数据集合，例如，假设统计目的为希望通过账户号码来统计某SNS某一天的登陆用户数，则云端服务器提取SNS操作日志中某一天所有访问过该SNS的用户账号的记录为待统计数据集合，其中用户账号作为待统计字段。

[0030] 在步骤S101，针对该待统计数据集合内的每条记录，分别根据该每条记录获取待统计字段的的索引号。

[0031] 在一个实施例中，上述的用户账号为无符号整形数，可直接将该用户账号字段的值作为其索引号。

[0032] 在步骤S102，获取位图，该位图包括多个位图单元，每个位图单元具有唯一的索引号。

[0033] 如图2所示，位图其实就是一个在内存内的位序列，每一位(单位位图)或者连续的多位作为一个位图单元(多位位图)。可以理解，基于二进制的原理，一个位图单元所能记录的值(或者说状态)为2^K种，其中K为位图单元的位数。以单位位图为例，每个位图单元的值

可为0或者1。

[0034] 获取上述位图过程可具体包括：创建该位图。该位图中位图单元的数量不少于该待统计字段可能值的数量。以待统计字段为4字节无符号整形数为例，则最大可能值为4294967196，则该位图至少应包括4294967196个位图单元。

[0035] 在步骤S103，与该待统计字段的索引号对应的位图单元内写入预定值。

[0036] 每两个相邻的位图单元之间的距离可定义为1。则每个位图单元与位图起始之间的距离即可为该位图单元的索引号。若位图的起始地址为S，则每个位图单元的地址为S+a*n，其中a为位图单元的索引号，n为位图单元内位的数目，若为单位位图，则地址可简化为S+n。

[0037] 上述对应的位图单元在该位图中的索引号例如可与该待统计字段的索引号相同。因此，步骤S103的写入过程实际上可根据上述的地址计算出位图单元的地址，直接通过位操作在该位图单元内写入该预定值。

[0038] 本实施中，需要统计的是唯一值的数量，因此采用0或者1标记该值出现过即可。但应注意的是，若采用0标记某个值出现过，则在使用位图前需要将所有位图单元的值初始化为1，若采用1标记某值出现过，则在使用位图前需要将所有位图单元的值初始化为0。

[0039] 在步骤S104，遍历该位图并统计值为该预定值的位图单元的个数。

[0040] 无论待统计数据集合内的对应字段的索引号相同的记录有多少条，与这些记录对应的位图单元内只有一个，且每次统计时存入的预定值相同，因此通过遍历该位图并统计值为该预定值的位图单元的个数，即可快速而高效地统计出待统计数据集合内的唯一值的数量。

[0041] 本发明提供的数据统计方法中，通过在位图内的索引号与待统计数据集合内的每条记录的对应字段的索引号相同的位图单元内写入预定值，来统计待统计数据集合中的唯一值，可有效减少内存开销，避免了内存溢出的风险，此外，还可避免哈希比较，从而提高数据统计的统计效率。

[0042] 第二实施例提供一种数据统计方法，其与第一实施例的方法相似其不同之处在于，步骤S101具体包括以下步骤：

[0043] 步骤S201，判断该待统计字段的值的类型是否为无符号整形；若是则执行步骤S202，否则执行步骤S203。

[0044] 步骤S202，将该待统计字段的值作为该待统计字段的索引号。

[0045] 步骤S203，该待统计字段的值按一一对应的方式映射为无符号整形数，并将映射得到的无符号整形数作为该待统计字段的索引号。

[0046] 例如，可采用哈希算法将待统计字段的值映射为无符号整形数，并将映射得到的无符号整形数作为该待统计字段的索引号。

[0047] 可以理解，待统计字段的值并不都是类似于第一实施例中的无符号整形的用户账号，其还可能是邮箱地址、IP地址、机器物理编号等字符串类型的值。而这些值无法直接映射至位图中，因此需要先将其转换为无符号整形数理，再根据得到的无符号整形数将其映射至位图中。

[0048] 本实施例的方法中，通过将非无符号整形数映射为无符号整形数，可以让本实施例的方法可以实现各种类型字段唯一值数量的统计。

[0049] 第三实施例提供一种数据统计方法，其与第一实施例的方法相似，其不同之处在于，如图4所示，步骤S102具体包括以下步骤：

[0050] 步骤S301，根据该待统计字段的索引号(N)及预定的分段长度(L)计算对应的分段索引号k＝N/L，并对计算结果取整；

[0051] 步骤S302，判断分段索引号为k的位图分段是否已被创建，若是，则执行步骤S303，否则执行步骤S304；

[0052] 步骤S303，获取分段索引号为k的位图分段；

[0053] 步骤S304，在预定的位图空间内创建一个新的位图分段，标记该新的位图的分段索引号为k，并记录其创建顺序索引号。

[0054] 参阅图5，其为按上述的方式建立的分段位图的示意图。图5所示的位图分段的创建顺序索引号依次从0到n。每个位图分段的长度为预定的长度L。位图段0的起始地址为S_0，结束地址为E_0，则每个位图段的起始地址为$S_n＝E_{n-1}+1＝S_0+L*n$。

[0055] 步骤S103可具体包括以下步骤：

[0056] 步骤S305，根据该待统计字段的索引号及(N)及预定的分段长度(L)计算偏移值offset＝N%L；

[0057] 步骤S306，将分段索引号为k的位图分段中与offset对应的位图单元内写入预定值。

[0058] 具体地，步骤S306中，根据以下公式获取与该待统计字段的索引号对应的位图单元的地址：$S_0+L*n+offset$，其中S_0为该位图空间的起始地址，n为该分段的创建顺序索引号；以及根据获取的地址写入预定值。

[0059] 可以理解，在图5所示的实例中，位图分段的创建顺序索引号从0开始，但可以理解，本实施例并不受此限制，例如，位图分段的创建顺序索引号可以从1或者其他任意整数开始，在计算地址时进行对应的转换即可。

[0060] 本实施例的方法中，将位图分段处理，只有当某个位图分段需要使用到时才创建该位图分段，因此，当待统计的字段的值没有覆盖到部分位图分段内时，这部分分段不会被创建，因此位图总体占用的内存空间得以进一步减少。相应地，由于位图空间的减少，遍历以统计数量的效率得以进一步提高。

[0061] 为进一步说明以上各实施例提供的数据统计方法与现有的唯一值统计技术在统计效率上的差异，申请人在对某SNS操作日志中的10.5亿流水进行数十次逐步递增测试。如图6所示，共计进行50次测试，原始记录数从2700万增加到13.5亿，独立用户数(唯一账户数)从1300万增加到1.46亿，每个账户为4字节正整数，最大账户一直为4294967196，最小账户一直为0，测试语言为Java，运行环境为JDK1.6.0_23，64位服务器，测试服务器的内存(RAM)为32G(吉)。

[0062] 一并参阅图7及图8，测试结果如下：

[0063] 1.采用HashSet方式，Java虚拟机(Java Virtual Machin，JVM)设置内存池大小参数(Xmx参数)为8G。1300万唯一用户数时，内存消耗已达到723M，之后快速增长到7798.2M；统计时间从14.8秒快速增长到20935.8秒。

[0064] 2.采用整段位图方式，位图总长度为最大无符号整形(unsigned int)，即4294967296。内存消耗一直为512M，统计时间从12秒增加到97.6秒。

[0065] 3.采用分段位图方式：

[0066] 1)段长设置为16K时，内存消耗从274.9M增加到291.6M，统计时间从12.2秒增加到306.7秒。

[0067] 2)段长设置为1024K时，内存消耗从333.2M增加到398.6M，统计时间从10.6秒增加到272.2秒。

[0068] 注：上述统计时间已剔除单纯遍历测试数据的时间。

[0069] 从以上测试结果中可以看出：

[0070] 一、以上各实施例提供的数据统计方法可解决现有的唯一值统计技术中的内存溢出问题。

[0071] 二、采用整段位图的方式，相比于分段位图的方式，消耗较多的内存空间，但统计所需要的时间较少。

[0072] 三、分段大小的不同会影响统计效率，可以理解，分段越小，越能节省内存空间，但相应统计效率(总体效率)会较低，分段越大，在统计效率上越接近整段位图的方式，但对存储空间的节省优胜有限，在实际操作中，选择1024K左右的分段大小可以取得一个相对平衡的值。

[0073] 此外，本实施例的采用的数据分布相当均匀，基于分段位图的原理，若数据越小或者越集中，分段位图的作用越明显，特别是在大数据的唯一值统计中有明显的优势。

[0074] 参阅图9，第四实施例提供一种数据统计装置，其包括数据获取模块41、索引号获取模块42、位图获取模块43、写入模块44、以及统计模块45。

[0075] 数据获取模块41用于获取待统计数据集合。

[0076] 索引号获取模块42用于针对数据获取模块41获取的待统计数据集合内的每条记录，分别根据该每条记录获取待统计字段的的索引号。

[0077] 具体地，在一个实例中，索引号获取模块42用于：若该待统计字段为无符号整形则将该待统计字段的值作为该待统计字段的索引号，否则将该待统计字段的值按一一对应的方式映射为无符号整形数，并将映射得到的无符号整形数作为该待统计字段的索引号。

[0078] 位图获取模块43用于获取位图，该位图包括多个位图单元，每个位图单元具有唯一的索引号。

[0079] 在一个实例中，位图获取模块43用于：创建该位图，该位图中位图单元的数量不少于该待统计字段可能值的数量。

[0080] 在一个实例中，参阅图10，位图获取模块43包括：分段索引号获取单元431以及分段获取模块432。

[0081] 分段索引号获取单元431用于根据该待统计字段的索引号(N)及预定的分段长度(L)计算对应的分段索引号k＝N/L，并对计算结果取整。

[0082] 分段获取模块432用于判断分段索引号为k的位图分段是否已被创建，若是，则获取分段索引号为k的位图分段；若否，则在预定的位图空间内创建一个新的位图分段，标记该新的位图的分段索引号为k，并记录其创建顺序索引号k。

[0083] 写入模块44用于在索引号与该每条记录的对应字段的索引号相同的位图单元内写入预定值。

[0084] 在一个实例中，参阅图11，写入模块44包括偏移值获取单元441、地址获取单元

442、以及写入单元443。

[0085] 偏移值获取单元441用于根据该待统计字段的索引号及(N)及预定的分段长度(L)计算偏移值offset＝N%L；

[0086] 地址获取单元442用于根据以下公式获取与该待统计字段的索引号对应的位图单元的地址:S0+L*k+offset,其中S0为该位图空间的起始地址;以及

[0087] 写入单元443用于根据获取的地址写入预定值。

[0088] 统计模块45用于遍历该位图并统计值为该预定值的位图单元的个数。

[0089] 关于本实施例的数据统计装置的其他细节还可参考前述各实施例的数据统计方法,在此不再赘述。

[0090] 根据上述的数据统计装置,通过在位图内的索引号与待统计数据集合内的每条记录的对应字段的索引号相同的位图单元内存入预定值写入预定值,来统计待统计数据集合中的唯一值,可有效减少内存开销,避免了内存溢出的风险,此外,还可避免哈希计算、存储与散列冲突解决、比较等,从而提高数据统计的统计效率。

[0091] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0092] 以上所述,仅是本发明的较佳实施例而已,并非对本发明作任何形式上的限制,虽然本发明已以较佳实施例揭露如上,然而并非用以限定本发明,任何熟悉本专业的技术人员,在不脱离本发明技术方案范围内,当可利用上述揭示的技术内容做出些许更动或修饰为等同变化的等效实施例,但凡是未脱离本发明技术方案内容,依据本发明的技术实质对以上实施例所作的任何简单修改、等同变化与修饰,均仍属于本发明技术方案的范围内。

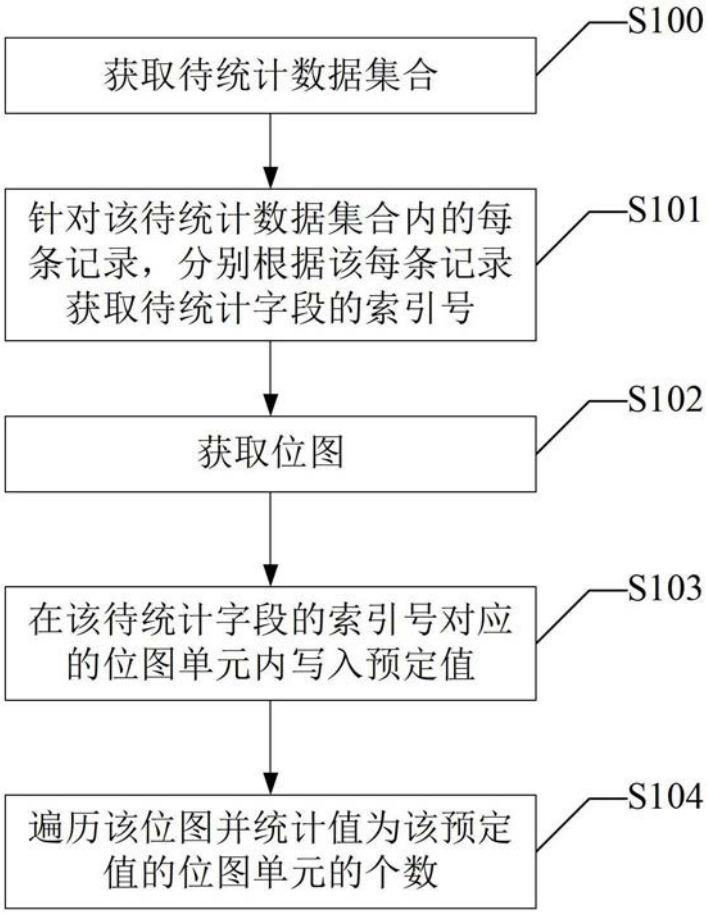

S100
获取待统计数据集合
S101
针对该待统计数据集合内的每条记录，分别根据该每条记录获取待统计字段的索引号
S102
获取位图
S103
在该待统计字段的索引号对应的位图单元内写入预定值
S104
遍历该位图并统计值为该预定值的位图单元的个数

图1

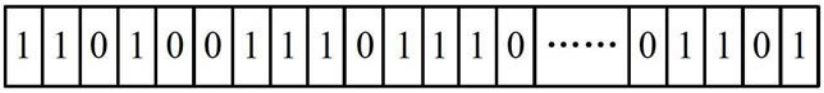

1 1 0 1 0 0 1 1 1 0 1 1 1 0 …… 0 1 1 0 1

图2

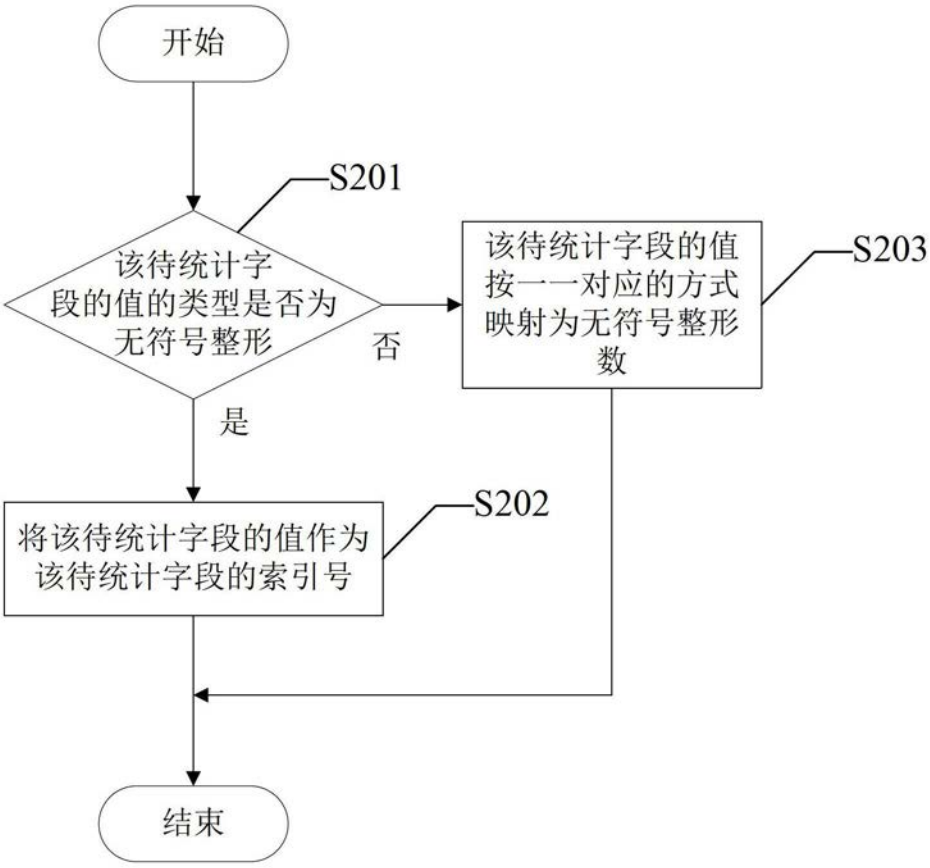

开始
S201
该待统计字段的值的类型是否为无符号整形
否
是
该待统计字段的值按一一对应的方式映射为无符号整形数
S203
将该待统计字段的值作为该待统计字段的索引号
S202
结束

图3

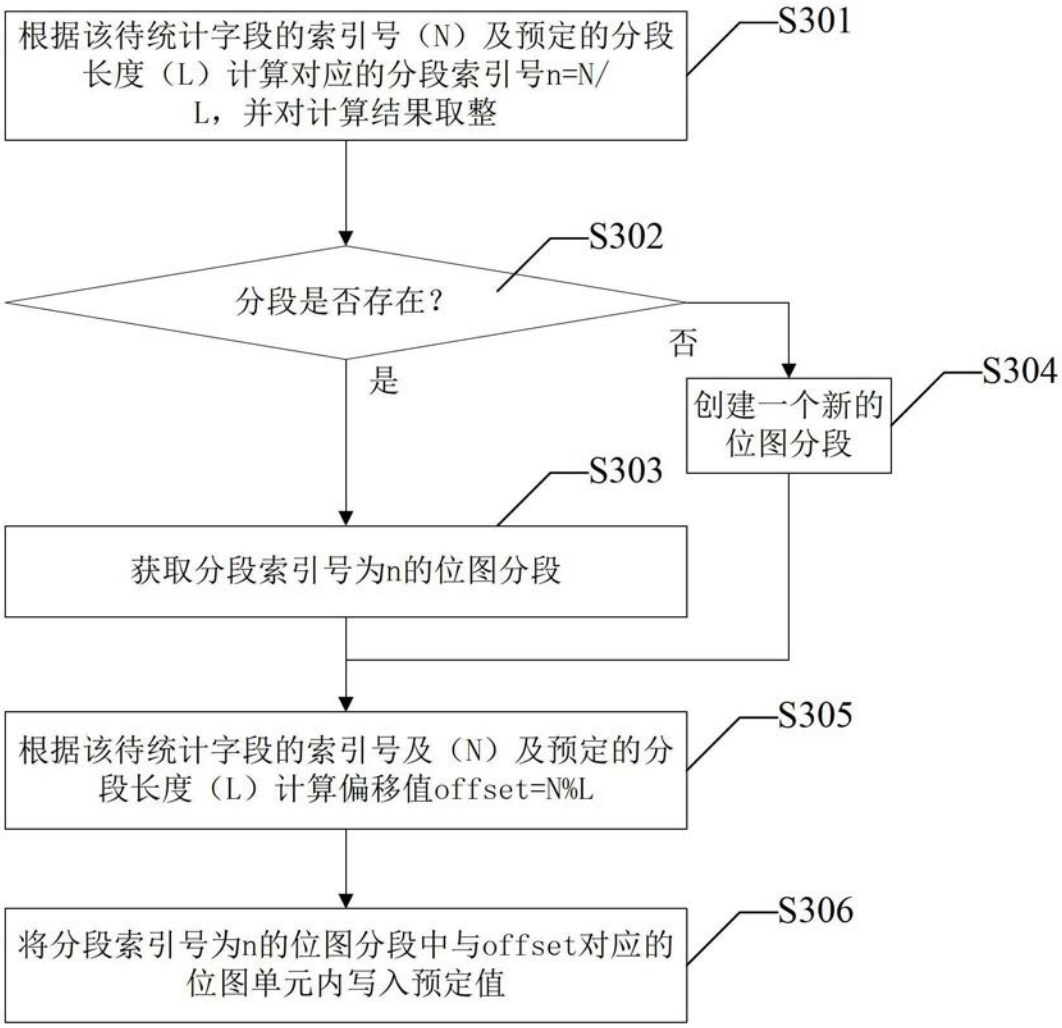

根据该待统计字段的索引号（N）及预定的分段长度（L）计算对应的分段索引号n=N/L，并对计算结果取整
S301
分段是否存在？
S302
否
是
创建一个新的位图分段
S304
S303
获取分段索引号为n的位图分段
根据该待统计字段的索引号及（N）及预定的分段长度（L）计算偏移值offset=N%L
S305
将分段索引号为n的位图分段中与offset对应的位图单元内写入预定值
S306

图4

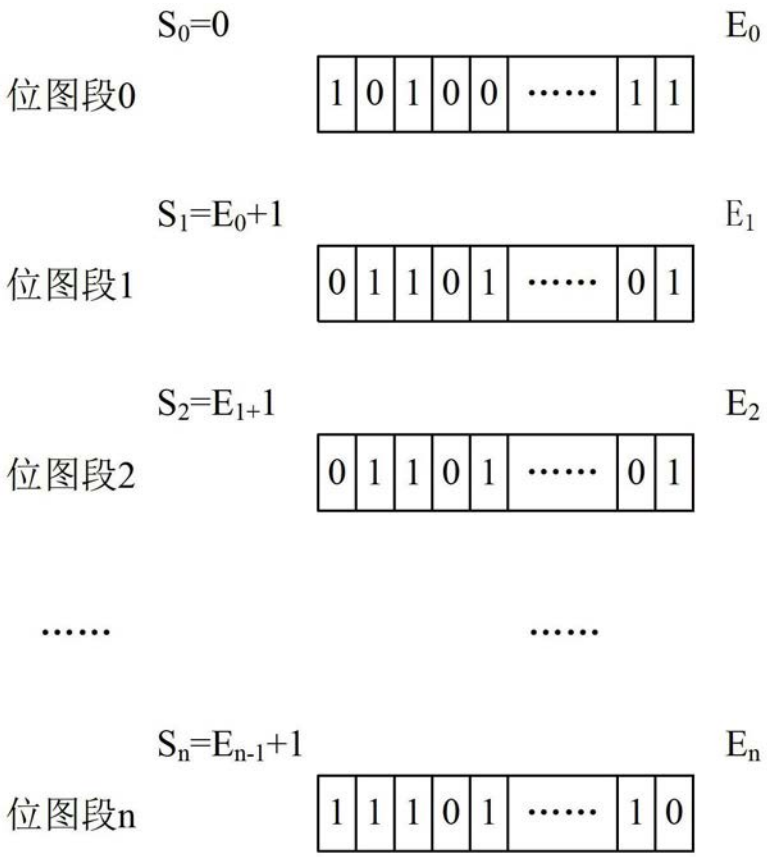

$S_0=0$
E_0
位图段0
1 0 1 0 0 …… 1 1
$S_1=E_0+1$
E_1
位图段1
0 1 1 0 1 …… 0 1
$S_2=E_{1+}1$
E_2
位图段2
0 1 1 0 1 …… 0 1
……
……
$S_n=E_{n-1}+1$
E_n
位图段n
1 1 1 0 1 …… 1 0

图5

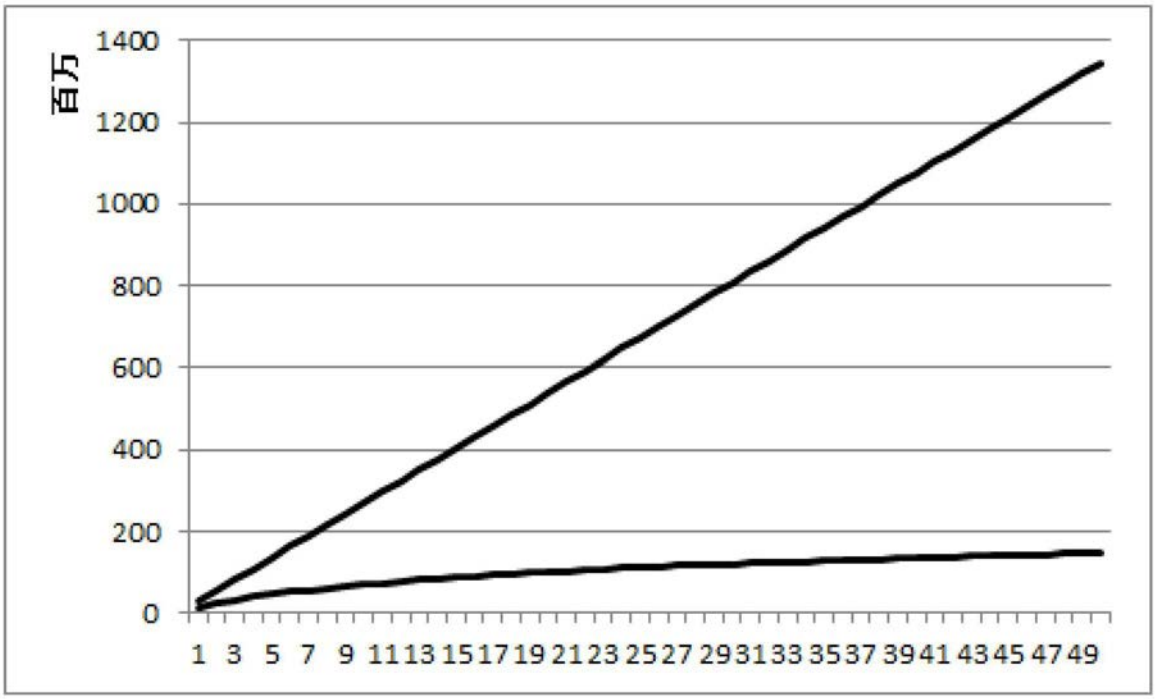

百万
1400
1200
1000
800
600
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49

图6

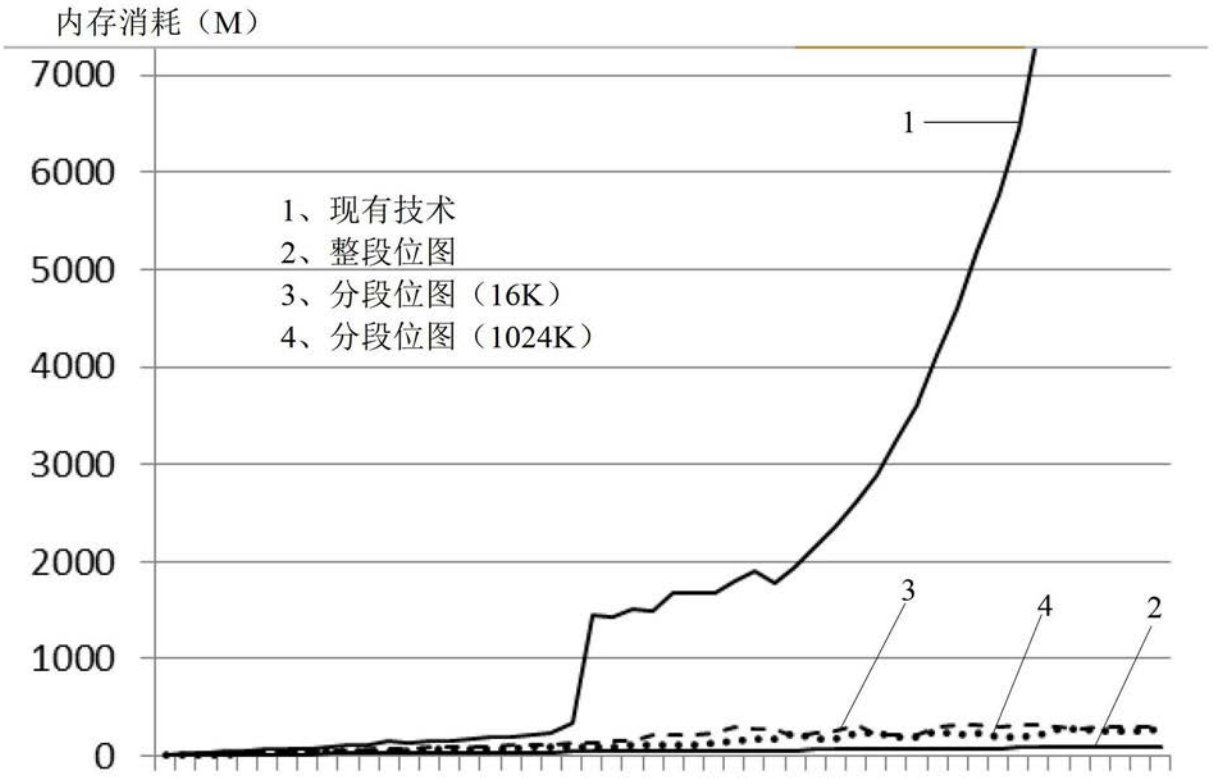

内存消耗（M）
7000
6000
5000
4000
3000
2000
1000
0
1、现有技术
2、整段位图
3、分段位图（16K）
4、分段位图（1024K）
1
3
4
2

图7

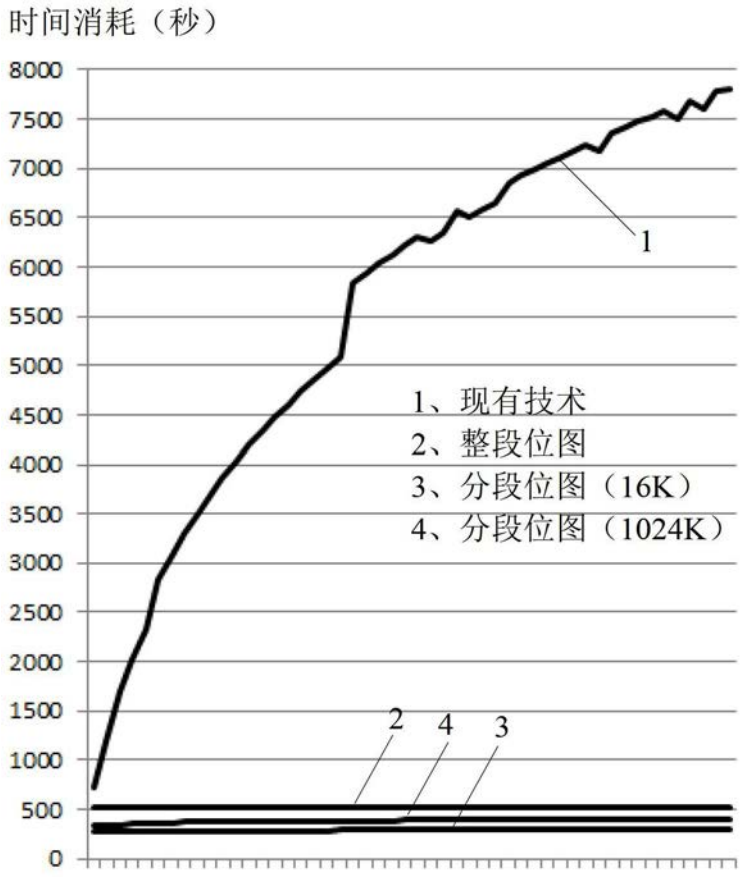

时间消耗（秒）
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
4
3
1、现有技术
2、整段位图
3、分段位图（16K）
4、分段位图（1024K）

图8

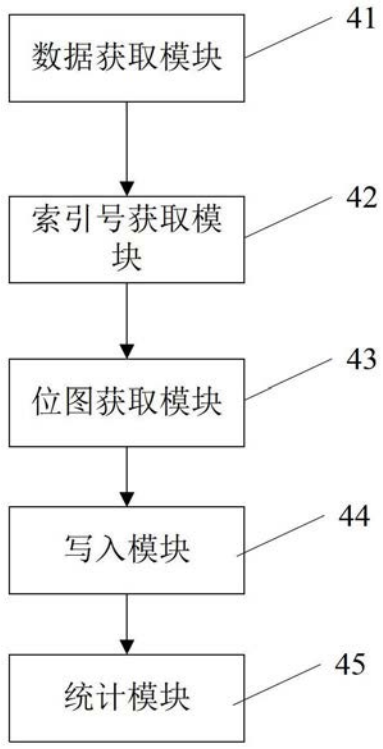

数据获取模块
41
索引号获取模块
42
位图获取模块
43
写入模块
44
统计模块
45

图9

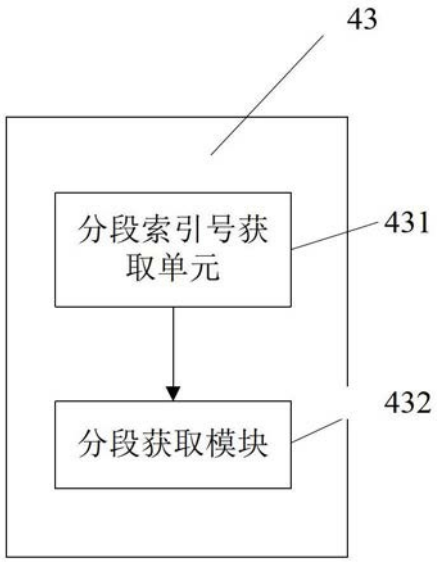

43
分段索引号获取单元
431
分段获取模块
432

图10

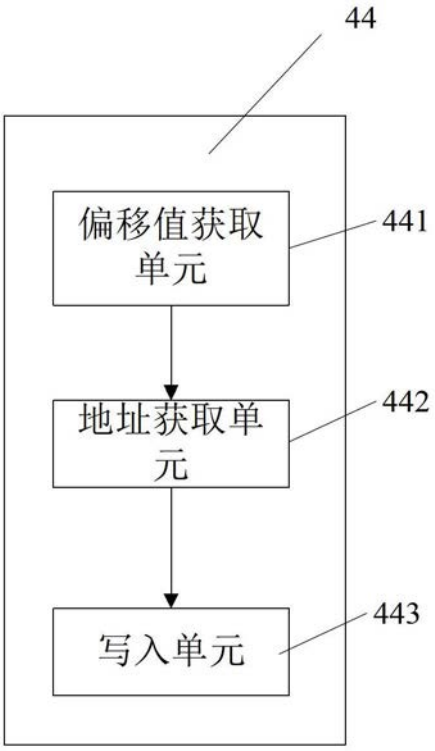


图11