(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 110297836 B
(45)授权公告日 2021.07.20

(21)申请号 201910622764.7
(22)申请日 2019.07.11
(65)同一申请的已公布的文献号
申请公布号 CN 110297836 A
(43)申请公布日 2019.10.01
(73)专利权人 杭州云梯科技有限公司
地址 310011 浙江省杭州市西湖区丰潭路669号新时代互联广场A座3013室
(72)发明人 田爽 陈立 施朝伟
(74)专利代理机构 成都九鼎天元知识产权代理有限公司 51214
代理人 阳佑虹
(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/2455(2019.01)
G06F 16/2457(2019.01)

审查员 杨琦

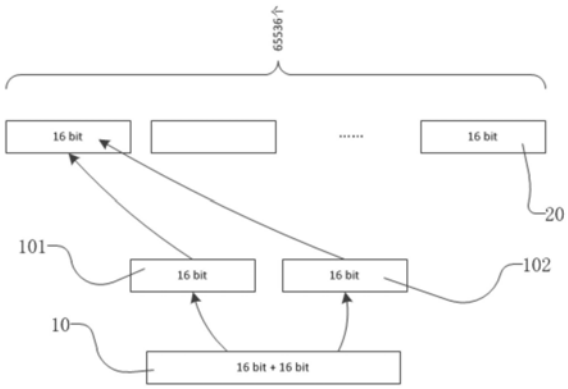
权利要求书1页 说明书5页 附图2页

(54)发明名称

基于压缩位图方式的用户标签存储方法和检索方法

(57)摘要

本发明公开了一种基于压缩位图方式的用户标签存储方法和检索方法。存储方法包括：将用户标签下的数据集对应的位图划分为若干成对的数据段；将存储空间划分为若干存储单元；每一对数据段的第一数据段均唯一关联有对应的存储单元；分别将每对数据段中的第二数据段存储到对应第一数据段所关联的存储单元，存储形式以数据类型和数据的数量为准。检索方法包括：划分待检索标签数据，通过第一字段查找存储空间，判断存储空间内是否存在匹配第二字段的记录。本发明存储方法较传统方式可大幅减小对存储空间的需求，便于对记录的提取和运算。智能存储方法可以确保对于存储空间的最小消耗。本发明的检索方法检索效率高。



CN 110297836 B

1.一种基于压缩位图方式的用户标签存储方法,其特征在于,包括:

将用户标签下的数据集对应的位图划分为若干成对的数据段;对应于用户标签下的每一条数据,每对数据段均包括第一数据段和第二数据段;

将存储空间划分为若干存储单元;

每一对数据段的第一数据段均唯一关联有对应的存储单元;

分别将每对数据段中的第二数据段存储到对应的第一数据段所关联的存储单元;所述存储单元至少能存储待存入的任一第二数据段;

所述分别将每对数据段中的第二数据段存储到对应第一数据段所关联的存储单元具体为:

判断待存入存储单元中的记录是否为连续形式,若否,则执行步骤A,否则执行步骤B;

A.若待存入存储单元中的记录的数量小于预定数量,则存储单元直接存储待存入的数据;否则,采用位图法存储待存入的数据;所述预定数量为存储单元存储数据的长度所能存储的用户标签类型的数据的数量;

B.对于连续形式的数据,存储单元关联存储初始记录和连续处理的次数。

2.如权利要求1所述的基于压缩位图方式的用户标签存储方法,其特征在于,每个所述数据段被划分的长度相同。

3.如权利要求1所述的基于压缩位图方式的用户标签存储方法,其特征在于,所述用户标签下的数据集对应的位图被划分的数据段对数满足2的正整数次幂。

4.如权利要求1-3之一所述的基于压缩位图方式的用户标签存储方法,其特征在于,所述用户标签下的数据集对应的位图被划分为两个数据段。

5.如权利要求1所述的基于压缩位图方式的用户标签存储方法,其特征在于,所述预定数量为2^k/16,k为存储单元存储数据的长度。

6.一种对权利要求1~5任一所述的基于压缩位图方式的用户标签存储方法存储的用户标签的检索方法,其特征在于,包括:

将待检索的用户标签数据划分为若干对数据段,所划分的数据段与权利要求1-5之一的用户标签存储方法中对用户标签所划分的数据段相同;待检索用户标签所划分的每对数据段均包括第一字段和第二字段,所述第一字段与第一数据段相应,所述第二字段与第二数据段相应;

对每一对第一字段和第二字段,均执行操作C-D:

C.通过第一字段,在存储空间中查找出相同第一数据段所关联的存储空间;

D.判断所述存储空间中是否存在与第二字段相同的第二数据段;

在对每一对第一字段和第二字段的操作结果均为存在对应的第二数据段时,则判定存储空间中存在与所述待检索的用户标签相匹配的记录,否则判定存储空间中不存在与所述待检索的用户标签相匹配的记录。

7.如权利要求6所述的对用户标签的检索方法,其特征在于,对每一对第一字段和第二字段执行操作C-D的步骤为:从第一对第一字段和第二字段开始,在对一对第一字段和第二字段执行操作C-D的操作结果为存在对应的第二数据段时,再执行对下一对第一字段和第二字段的操作,若任一对第一字段和第二字段的操作结果为不存在对应的第二数据段时,则判定存储空间中不存在与所述待检索的用户标签相匹配的记录。

基于压缩位图方式的用户标签存储方法和检索方法

技术领域

[0001] 本发明涉及数据存储领域，尤其是一种利用压缩位图以将数据进行分块存储的方式来存储用户标签下的数据集的方法，以及基于该存储方法的用户标签检索方法。

背景技术

[0002] 为实现个性化营销，精准推送，用户画像等需求，现在许多互联网公司都开发了用户标签系统，即对不同的用户打上各自的标签，构造出属于每个用户的独一无二的用户画像，但在互联网公司普遍用户量上千万，标签数成千上万的情况下，用户标签系统的设计、存储就显得至关重要，若不能对用户标签的相关数据进行实时快速的查询、运算，则不能顺利的支撑相关业务的发展。

[0003] 假设一个标签对应10000000个用户，若采用常用的关系型数据库将每个标签对应每个用户id存储为一条记录，那么单个标签的用户数据就会占用10000000条记录，若需要存储1000个标签对应的用户数据，平均每个标签对应10000000用户，那么关系型数据库单表需要存储10000000000行的数据，明显超过了关系型数据库的存储极限，也不符合关系型数据库设计推荐的数据量。

[0004] 再假设存在A、B两个标签描述数据，如果需要求出同时具有A标签及B标签的用户，则需要在数据库层计算A标签及B标签用户数据的交集，那么对于标签用户表，需要做SELFJOIN，然后再过滤掉不符合条件的行，因为JOIN第一步就是对表记录进行笛卡尔积，若标签用户表数据量为n，那么第一步进行笛卡尔积后的数据量将为n的平方，因为标签用户表数据量本来就很大了，再进行笛卡尔积操作，会大量占用关系型数据库的计算资源，很可能会造成关系型数据库的CPU占用率过高而导致不可用。

发明内容

[0005] 本发明的发明目的在于：针对上述存在的问题，提供一种利用压缩位图的方式，分块存储用户标签的方法。以提高对存储空间的利用效率，提高对数据操作(查找、计算)的便捷性。

[0006] 本发明采用的技术方案如下：

[0007] 一种基于压缩位图方式的用户标签存储方法，其包括：

[0008] 将用户标签下的数据集对应的位图(即采用位图存储用户标签下的数据集的结果)划分为若干成对的数据段，对应于用户标签下的每一条数据(即每一个用户的标签数据)，每对数据段均包括第一数据段和第二数据段；即对应于被划分的位图，每一条数据均被划分为若干对数据段，每对数据段均包括第一数据段和第二数据段；

[0009] 将存储空间划分为若干存储单元；

[0010] 每一对数据段的第一数据段均唯一关联有对应的存储单元；

[0011] 分别将每对数据段中的第二数据段存储到对应的第一数据段所关联的存储单元；该存储单元至少能存储待存入的任一第二数据段，即存储单元存储数据的长度不低于待存

入的任一条数据的数据长度。

[0012] 上述方法,将用户标签下的数据集对应的位图(全长度的位图)划分为若干段落,将部分段落作为存储索引,关联于对应的存储单元,部分段落作为存储的数据的形式进行存储。该方式可在现有的位图存储基础上,大幅缩减了对存储空间的需求,提高了存储空间的利用率。用户越多,效果越明显。

[0013] 进一步的,每个数据段被划分的长度相同。

[0014] 进一步的,所述用户标签下的数据集对应的位图被划分的数据段对数满足2的正整数次幂。上述配置均是便于计算机的处理。

[0015] 进一步的,所述用户标签下的数据集对应的位图被划分为两个数据段。

[0016] 划分为两个数据段足以满足绝大多数场景的需求,兼顾了存储空间和存储/检索效率的需求。

[0017] 进一步的,上述分别将每对数据段中的第二数据段存储到对应第一数据段所关联的存储单元具体为:

[0018] 判断待存入存储单元中的记录是否为连续形式,若否,则执行步骤A,否则执行步骤B;

[0019] A.若待存入存储单元中的记录的数量小于预定数量,则存储单元直接存储待存入的数据,即对用户标签下每一条数据对应于位图的第二数据段以数据原本的类型进行存储;否则,采用位图法存储待存入的数据,即采用位图法存储用户标签下的数据集对应的位图的第二数据段;所述预定数量为存储单元存储数据的长度所能存储的用户标签类型的数据的数量;

[0020] B.对于连续形式的数据,存储单元关联存储初始记录和连续处理的次数。

[0021] 采用上述方式,可以确保存储过程中存储空间消耗最低,且属于动态自调整的形式。

[0022] 进一步的,上述预定数量为2^k/16,k为存储单元存储数据的长度。即针对于整数类型的数据。

[0023] 本发明提供了一种用户标签检索方法,其包括:

[0024] 将待检索的用户标签数据划分为若干对数据段,所划分的数据段与上述用户标签存储方法中对用户标签下的数据集对应的位图所划分的数据段相同(即划分的段数、每段对应的长度均相同);待检索用户标签数据所划分的每对数据段均包括第一字段和第二字段,所述第一字段与第一数据段相应,所述第二字段与第二数据段相应;

[0025] 对每一对第一字段和第二字段,均执行操作A—B:

[0026] A.通过第一字段,在存储空间中查找出相同第一数据段所关联的存储空间;

[0027] B.判断所述存储空间中是否存在与第二字段相同的第二数据段;

[0028] 在对每一对第一字段和第二字段的操作结果均为存在对应的第二数据段时,则判定存储空间中存在与所述待检索的用户标签相匹配的记录,否则判定存储空间中不存在与所述待检索的用户标签相匹配的记录。

[0029] 上述方式,无需对完整的用户标签进行逐位对比,通过快速定位存储单元的方式,同时,仅需对比部分位上的数据,极大地提高了检索效率。同时,多个存储单元可并行检索,检索速度快。

[0030] 进一步的，对每一对第一字段和第二字段执行操作A－B的步骤为：从第一对第一字段和第二字段开始，在对一对第一字段和第二字段执行操作A－B的操作结果为存在对应的第二数据段时，再执行对下一对第一字段和第二字段的操作，若任一对第一字段和第二字段的操作结果为不存在对应的第二数据段时，则判定存储空间中不存在与所述待检索的用户标签相匹配的记录。

[0031] 顺序判定的方式在判定存在不匹配记录时即跳出检索，可以节省后续不必要的检索运算，减少了运算消耗和检索时间。

[0032] 综上所述，由于采用了上述技术方案，本发明的有益效果是：

[0033] 1、在存储方面，传统关系数据库每个标签对应每个用户需要一条记录进行保存，无法满足互联网公司大用户量的存储，若拆分至多表存储，则无法使用数据库原生的JOIN进行多标签共同用户的计算等操作。本发明的存储方法大幅减小了对存储空间的需求。通过计算可知，即使用户量达到上亿级，单个用户标签数据所需的存储空间也不到10MB。智能化的存储方式方式可以使得对数据的存储所消耗的存储空间最少。

[0034] 2、在计算方面，传统关系数据库的存储方式若需要计算多标签的交集、并集的用户数据，需要将原表进行JOIN然后过滤掉不符合条件的记录，此处需要进行笛卡尔积运算，所消耗的资源随着标签数、标签对应用户数的增长而增长，不能满足大规模的交集、并集等运算，关系型数据库的设计只支持垂直扩容，单机能达到的性能有限。本发明的存储方法采用1G内存的单机即可支持数百个标签的实时运算，一台普通的PC机即可满足大多数场景下的计算要求，且可根据需求灵活配置。

[0035] 3、本发明的检索方法可对待检索的数据进行多点并行检索，检索效率高，且一旦发现不匹配时即终止检索，节省了不必要的运算。

附图说明

[0036] 本发明将通过例子并参照附图的方式说明，其中：

[0037] 图1是32位用户标签存储示意图。

[0038] 图2是检索方法的一个实施例。

[0039] 图中，10为用户标签下的数据集对应的位图，101为位图的高16位，102为位图的低16位，20为存储单元。

具体实施方式

[0040] 本说明书中公开的所有特征，或公开的所有方法或过程中的步骤，除了互相排斥的特征和/或步骤以外，均可以以任何方式组合。

[0041] 本说明书(包括任何附加权利要求、摘要)中公开的任一特征，除非特别叙述，均可被其他等效或具有类似目的的替代特征加以替换。即，除非特别叙述，每个特征只是一系列等效或类似特征中的一个例子而已。

[0042] 实施例一

[0043] 一种基于压缩位图方式的用户标签存储方法，其包括：

[0044] 将用户标签下的数据集对应的位图划分为若干成对的数据段。每对数据段中，一个作为索引，另一个作为检索目标。每一段的数据长度均满足2^n(n为正整数)位。在一个实

施例中，数据段的长度均分。对于位图而言，其以位图法存储有若干条用户标签数据，对于位图段落的划分，会对其下每一条用户标签数据进行对应的划分，即以位图形式被存储的每一条用户标签数据被划分为若干对数据段。

[0045] 将存储空间划分为若干个存储单元，每一存储单元的存储空间均至少能存储待存入的检索目标。在一个实施例中，所有存储单元的存储空间均相同。

[0046] 分别将每对数据段中的检索目标存储到与之成对的索引所唯一关联的存储单元中，每一第一数据段均关联有对应的存储单元。即将每对数据段中的索引作为对应检索目标所存储到的存储单元的检索路径(例如作为存储单元的编号)。通过存储单元的检索路径及其内所存储的数据(检索目标)，即可完成对完整用户标签的存储。传统的存储方式会将用户标签下的数据集进行完整的存储，其需要适配于用户标签数据长度的数据存储空间，在所存储的用户标签数量较少时，则会对存储空间造成极大的浪费，同时，由于所存储的数据长度较长(对应于用户标签的最大值)，在查找、计算时也极为不便。本实施例将用户标签划分为若干部分进行存储，通过非完全存储的方式实现了对完整用户标签的存储，所需存储空间小。该种存储方式便于检索和(位)计算。

[0047] 实施例二

[0048] 对于对位图划分的段落和长度，是出于所需存储空间的数量(用户数)和容量的综合考虑。本实施例以将位图平均划分为两段(即一个段落对)为例。如图1所示，以32位位图为例，位图被划分为高16位和低16位，将高16位作为索引，低16位作为检索目标。32位无符号数据存储用户标签下的数据集的最大容量为2^32条记录，即最大支持超过42.9亿个记录的存储，显然已经满足现有需求。

[0049] 实施例三

[0050] 本实施例公开了上述将检索目标存入对应存储单元的具体方法，其包括：

[0051] 判断待存入存储单元中的记录(即用户标签下的数据集中，待存入该存储单元的数据)是否为连续形式，若否，则执行步骤A，否则执行步骤B；

[0052] A:若待存入存储单元中的记录的数量小于预定数量(该预定数量在数值上与存储单元的长度有关，对应于用户标签类型为整型(占2个字节)的数据，预定数量＝2^k/16，k为存储单元的长度(位数)，其余类型的用户标签数据同理)，则存储单元直接存储用户标签下每一条数据对应于位图的第二数据段的数据；否则，采用位图方式存储用户标签下的数据集对应的位图的第二数据段。这是出于对标签的检索效率和存储空间利用效率的考虑。例如，对应于存储单元的长度为16位、32位位图平均分为两段数据段的情况，在用户标签下的数据集小于4096条时，存储单元直接将用户标签下的数据以整数形式存储，由于每个整数(用户标签数据的低16位)占2B，则所占用的空间小于4096＊2B＝8192B；若用于标签下的数据集达到4096条，则存储空间采用位图方式存储位图第二数据段的数据，即无论数据集有多少条，都会占用2^16bit＝8KB的存储空间。

[0053] B.对于连续形式的数据，采用连续值压缩存储方式，即关联存储初始值和连续处理的次数。例如对于[1,1000]这样的数据，存储形式为“1,999”，表示从1开始，后面连续999次重复处理(递增)，表示后续存在999个数值(记录)。该方式能将存储空间从2000B(即直接存储为整数形式的方式，2B＊1000)，或者8KB(位图存储方式)降低至4(即2B＊2)个字节。

[0054] 由此可见，对于不同的存储量需求而言，采用的存储方式对所需的存储空间有较

大的影响。存储单元存储数据的方式根据计划存入的数据的数量及数据的分布情况进行动态调整，可以使得数据对存储单元占用的存储空间最小。

[0055] 基于实施例二中对位图的划分方式，对于存储介质而言，划分为2^16(即65536)个存储单元，每一存储单元在采用位图存储方式时需要的存储空间为2^16bit即8KB，采用直接存储数据的方式则由待存入的记录数量确定，为2B＊N，N为待存入的数据条数。从对记录的处理方面，对于CPU的L1Cache，根据不同处理器实现，均可同时载入多个存储单元缓存，提高运行速度，对于常用的bitCount操作，可直接利用CPU的popcnt/cnt指令在CPU层直接获取数据，避免在应用层进行相关计算，降低相关操作的耗时。从对存储空间的利用方面，假设在每个存储单元中存储的数据较少，以3条记录为例，若存储单元直接存储3条数据的低16位，则需要的存储空间仅为3＊(16/8)＝6B，若仍采用位图存储方式，则需要2＊16bit＝8KB的存储空间。

[0056] 实施例四

[0057] 本实施例公开了基于实施例一中存储方法的用户标签检索方法，其包括：

[0058] 以存储用户标签的形式，将待检索的用户标签划分为若干成对的数据段。每对数据段均包括第一字段和第二字段，第一字段对应于索引，第二字段对应于检索目标。

[0059] 对于每一对数据段，均执行以下操作：

[0060] 通过第一字段，在存储空间中查找出相同索引所对应的存储空间；

[0061] 判断该存储空间中是否存在与第二字段相同的检索目标。

[0062] 在对每一对数据段的操作结果均为存在对应的检索目标时，则表明存储空间中存在与待检索的用户标签相匹配的记录。

[0063] 考虑到检索效率，如图2所示，上述对每一对数据段的操作，为以此对每一对数据段执行操作，并在对一对数据段执行的操作结果为存在对应的检索目标时，再执行对下一对数据段的操作，否则，判定存储空间中不存在与待检索的用户标签匹配的记录。

[0064] 本发明并不局限于前述的具体实施方式。本发明扩展到任何在本说明书中披露的新特征或任何新的组合，以及披露的任一新的方法或过程的步骤或任何新的组合。

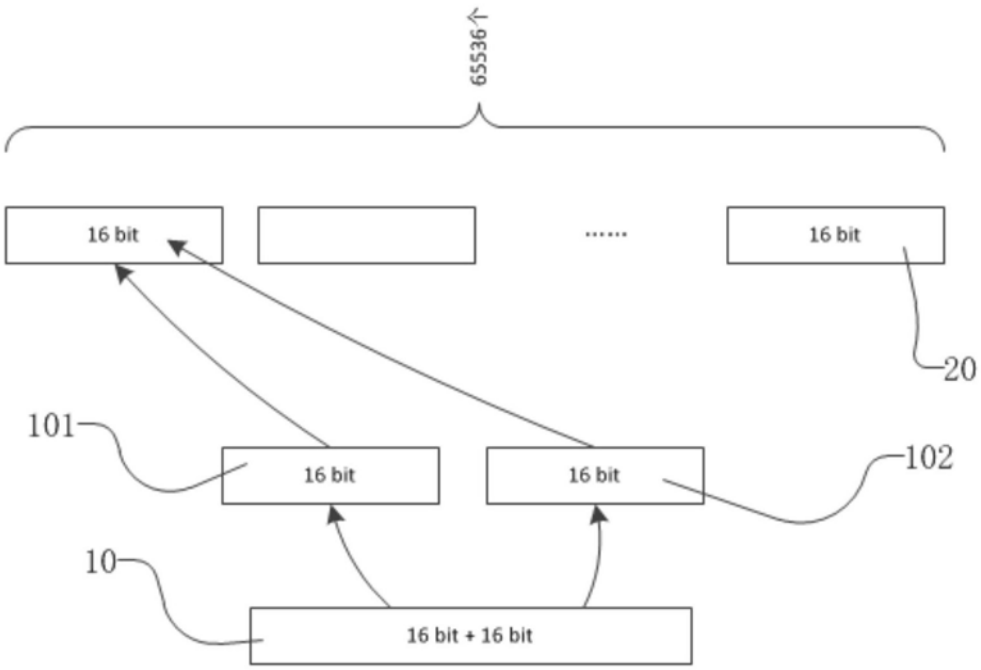


图1

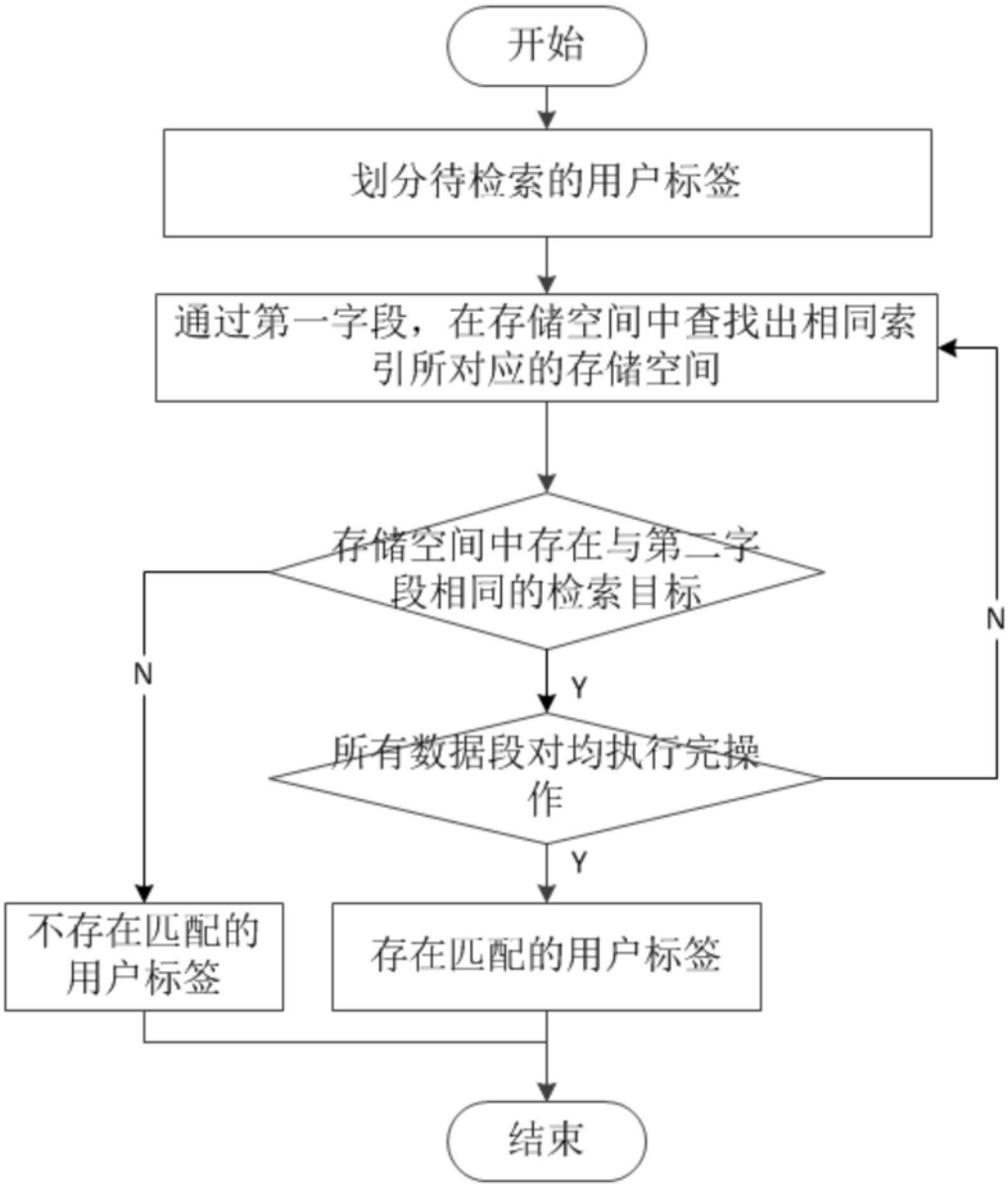

开始
划分待检索的用户标签
通过第一字段，在存储空间中查找出相同索引所对应的存储空间
存储空间中存在与第二字段相同的检索目标
N
Y
所有数据段对均执行完操作
N
Y
不存在匹配的用户标签
存在匹配的用户标签
结束

图2