(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 105630803 B
(45)授权公告日 2019.07.05

(21)申请号 201410599318.6

(22)申请日 2014.10.30

(65)同一申请的已公布的文献号
申请公布号 CN 105630803 A

(43)申请公布日 2016.06.01

(73)专利权人 国际商业机器公司
地址 美国纽约

(72)发明人 谢芳全 李峰 李起成 梅立军
李少春 陈昊

(74)专利代理机构 北京市中咨律师事务所 11247
代理人 于静 张亚非

(51)Int.Cl.
G06F 16/13(2019.01)

(56)对比文件
CN 102651007 A,2012.08.29,
CN 102521282 A,2012.06.27,
US 2014229427 A1,2014.08.14,
CN 102262640 A,2011.11.30,
US 2009094236 A1,2009.04.09,

审查员 单娟

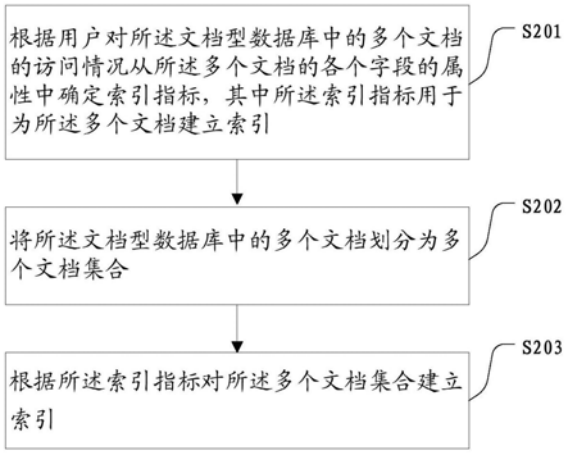
权利要求书3页 说明书10页 附图6页

(54)发明名称

文档型数据库建立索引的方法和装置

(57)摘要

本发明涉及数据库,其公开一种文档型数据库建立索引的方法,包括:根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标,其中所述索引指标用于为所述多个文档建立索引;将所述多个文档划分为多个文档集合;根据所述索引指标对所述多个文档集合建立索引。根据本发明实施例的方法利用文档型数据库易于分块的特点将文档型数据库中的多个文档划分为文档集合,针对文档集合能够有效地实现为文档型数据库建立索引。



CN 105630803 B

1.一种文档型数据库建立索引的方法,包括:

根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标,其中所述索引指标用于为所述多个文档建立索引;将所述多个文档划分为多个文档集合;

根据所述索引指标对所述多个文档集合建立索引,

其中根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标包括:

记录一段时间内用户对所述多个文档的各个字段的操作;

统计出所述多个文档的各个字段的属性并统计出针对每个属性的相同操作的次数;

根据每个属性的相同类型的操作次数从所述属性中确定索引指标。

2.根据权利要求1所述的方法,其中响应于确定所述各个字段中的字段b的属性包括子属性,所述针对每个属性的相同操作的次数是针对所述字段b的属性的全部子属性的相同操作的次数总和。

3.根据权利要求1所述的方法,其中用户对所述文档型数据库中的多个文档的各个字段的操作包括以下操作中的至少一个:查询操作、写操作、分组和排序操作。

4.根据权利要求1所述的方法,其中所述多个文档集合的数目大于或者等于索引指标的数目。

5.根据权利要求1所述的方法,其中根据所述索引指标对所述多个文档集合建立索引包括:

按照所述索引指标的数目选取文档集合;

为选取的每个文档集合分配索引指标并按照分配的索引指标建立索引;

响应于确定存在已完成建立索引的文档集合,获取分配给所述完成建立索引的文档集合的索引指标a;

响应于确定存在未分配索引指标的文档集合,选取一个未分配索引指标的文档集合并按照索引指标a建立索引。

6.根据权利要求5所述的方法,其中根据所述索引指标对所述多个文档集合建立索引进一步包括:

响应于确定不存在未分配索引指标的文档集合,判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引;

响应于判断结果为否,选取一个未分配过索引指标a的文档集合并按照索引指标a建立索引。

7.根据权利要求5所述的方法,进一步包括:

获取新增索引指标;

响应于确定存在未分配索引指标的文档集合,选取一个未分配索引指标的文档集合并按照新增索引指标建立索引。

8.根据权利要求7所述的方法,进一步包括:

响应于确定不存在未分配索引指标的文档集合,判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引;

响应于判断结果为否,为每个文档集合统计分配的索引指标的数目;

为索引指标的分配数目最少的文档集合分配新增索引指标并按照新增索引指标建立索引。

9.根据权利要求5所述的方法，进一步包括：

获取废弃索引指标；

响应于确定存在已按所述废弃索引指标建立索引的文档集合，删除按照所述废弃索引指标建立的索引并且删除所述废弃索引指标。

10.一种文档型数据库建立索引的装置，包括：

第一确定模块，被配置为根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标，其中所述索引指标用于为所述多个文档建立索引；

划分模块，被配置为将所述多个文档划分为多个文档集合；

建立模块，被配置为根据所述索引指标对所述多个文档集合建立索引，

其中第一确定模块进一步包括：

记录模块，被配置为记录一段时间内用户对所述多个文档的各个字段的操作；第一统计模块，被配置为统计出所述多个文档的各个字段的属性并统计出针对每个属性的相同操作的次数；

第二确定模块，被配置为根据每个属性的相同类型的操作次数从所述属性中确定索引指标。

11.根据权利要求10所述的装置，其中响应于确定所述各个字段中的字段b的属性包括子属性，所述针对每个属性的相同操作的次数是针对所述字段b的属性的全部子属性的相同类型的操作次数的总和。

12.根据权利要求10所述的装置，其中用户对所述文档型数据库中的多个文档的各个字段的操作包括以下操作中的至少一个：查询操作、写操作、分组和排序操作。

13.根据权利要求10所述的装置，其中所述多个文档集合的数目大于或者等于索引指标的数目。

14.根据权利要求10所述的装置，其中建立模块包括：

第一选取模块，被配置为按照所述索引指标的数目选取文档集合；

第一分配模块，被配置为为选取的每个文档集合分配索引指标并按照分配的索引指标建立索引；

第一获取模块，被配置为响应于确定存在已完成建立索引的文档集合，获取分配给所述完成建立索引的文档集合的索引指标a；

第二选取模块，被配置为响应于确定存在未分配索引指标的文档集合，选取一个未分配索引指标的文档集合并按照索引指标a建立索引。

15.根据权利要求14所述的装置，其中建立模块进一步包括：

第一判断模块，被配置为响应于确定不存在未分配索引指标的文档集合，判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引；

第三选取模块，被配置为响应于判断结果为否，选取一个未分配过索引指标a的文档集合并按照索引指标a建立索引。

16.根据权利要求14所述的装置，进一步包括：

第二获取模块，被配置为获取新增索引指标；

第四选取模块，被配置为响应于确定存在未分配索引指标的文档集合，选取一个未分配索引指标的文档集合并按照新增索引指标建立索引。

17.根据权利要求16所述的装置，进一步包括：

第二判断模块，被配置为响应于确定不存在未分配索引指标的文档集合，判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引；

第二统计模块，被配置为响应于判断结果为否，为每个文档集合统计分配的索引指标的数目；

第二分配模块，被配置为为索引指标的分配数目最少的文档集合分配新增索引指标并按照新增索引指标建立索引。

18.根据权利要求14所述的装置，进一步包括：

第三获取模块，被配置为获取废弃索引指标；

删除模块，被配置为响应于确定存在已按所述废弃索引指标建立索引的文档集合，删除按照所述废弃索引指标建立的索引并且删除所述废弃索引指标。

文档型数据库建立索引的方法和装置

技术领域

[0001] 本发明涉及数据库,更具体地,涉及基于文件型数据库建立索引的方法和装置。

背景技术

[0002] 随着互联网Web 2.0的兴起,NoSQL非关系型数据库成为一个极其热门的新领域,面对数据库高并发读写的需求,对海量数据的高效率存储和访问的需求,对数据库的高可扩展性和高可用性的需求,关系型数据库已经力不从心。与关系型数据库相比,NoSQL数据库具有灵活的可扩展性,NoSQL数据库种类繁多,但是一个共同的特点都是去掉关系数据库的关系型特性,数据之间无关系,这样就非常容易扩展,在架构的层面上带来了可扩展的能力。文档型数据库是非关系型数据库中非常重要的一个分支,它主要用来存储、索引并管理面向文档的数据或者类似的半结构化数据。顾名思义,文档型数据库(面向文档数据库)的关键核心概念即文档(Document),它是数据库中最小的单位。MongoDB是目前最为流行的NoSQL数据库,它是一种面向集合、模式无关的文档型数据库。其中数据以“集合”的方式进行分组,每个集合都有单独的名称并可以包含无限数量的文档。这里的集合同关系型数据库中的表(table)类似,唯一的区别就是它并没有任何明确的模式(schema)。

[0003] 创建数据库索引是数据库管理的一个重要方面,数据库索引是对数据库表中一列或多列的值进行排序的一种数据结构,这些数据结构以某种方式引用(指向)数据,以协助快速查询、更新数据库表中数据。关系型数据库通常以表结构存储,索引的建立可简单的仅针对固定的某些字段。而文档型数据库通常是不限定字段结构的,且在文档型数据库使用过程中会不断的有新的文档引入新的字段结构,因此预先选取某些固定字段不能有效地应对文档型数据库文档字段的动态变化。此外,由于数据分块的难度较大,针对关系型数据库的索引建立是针对表中的全部数据,当数据量很大时,尤其针对在线提供服务的非关系型数据库文档中的全部数据建立索引,建立索引期间访问数据库的性能变得很差。

[0004] 因此,需要一种有效地为文档型数据库建立索引的方法。

发明内容

[0005] 根据本发明的一个方面,提供一种文档型数据库建立索引的方法,包括:根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标,其中所述索引指标用于为所述多个文档建立索引;将所述多个文档划分为多个文档集合;根据所述索引指标对所述多个文档集合建立索引。

[0006] 根据本发明的另一个方面,提供一种文档型数据库建立索引的装置,包括:第一确定模块,被配置为根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标,其中所述索引指标用于为所述多个文档建立索引;划分模块,被配置为将所述多个文档划分为多个文档集合;建立模块,被配置为根据所述索引指标对所述多个文档集合建立索引。

[0007] 根据本发明实施例的方法和装置利用文档型数据库易于分块的特点将文档型数

据库中的多个文档划分为文档集合，针对文档集合能够有效地实现为文档型数据库建立索引。

附图说明

[0008] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0009] 图1示出了适于用来实现本发明实施方式的示例性计算机系统/服务器12的框图。

[0010] 图2示出根据本发明实施例的一种文档型数据库建立索引的方法。

[0011] 图3示出根据本发明实施例的文档1、2和3的属性的树状结构。

[0012] 图4示出在图3的属性的树状结构上标识出针对每个属性的相同类型的操作次数。

[0013] 图5示出根据本发明的实施例根据所述索引指标对所述文档集合中的文档建立索引的流程。

[0014] 图6示出根据图5的流程根据新增索引指标建立索引的流程。

[0015] 图7示出根据图5的流程处理废弃索引指标的流程。

[0016] 图8示出根据本发明实施例的文档型数据库建立索引的装置。

具体实施方式

[0017] 下面将参照附图更详细地描述本公开的优选实施方式。虽然附图中显示了本公开的优选实施方式，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0018] 图1示出了适于用来实现本发明实施方式的示例性计算机系统/服务器12的框图。图1显示的计算机系统/服务器12仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0019] 如图1所示，计算机系统/服务器12以通用计算设备的形式表现。计算机系统/服务器12的组件可以包括但不限于：一个或者多个处理器或者处理单元16，系统存储器28，连接不同系统组件(包括系统存储器28和处理单元16)的总线18。

[0020] 总线18表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0021] 计算机系统/服务器12典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机系统/服务器12访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0022] 系统存储器28可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(RAM)30和/或高速缓存存储器32。计算机系统/服务器12可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统34可以用于读写不可移动的、非易失性磁介质(图1未显示，通常称为“硬盘驱动器”)。尽管图1中未示

出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线18相连。存储器28可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0023] 具有一组(至少一个)程序模块42的程序/实用工具40,可以存储在例如存储器28中,这样的程序模块42包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42通常执行本发明所描述的实施例中的功能和/或方法。

[0024] 计算机系统/服务器12也可以与一个或多个外部设备14(例如键盘、指向设备、显示器24等)通信,还可与一个或者多个使得用户能与该计算机系统/服务器12交互的设备通信,和/或与使得该计算机系统/服务器12能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22进行。并且,计算机系统/服务器12还可以通过网络适配器20与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器20通过总线18与计算机系统/服务器12的其它模块通信。应当明白,尽管图中未示出,可以结合计算机系统/服务器12使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0025] 图2示出根据本发明实施例的一种文档型数据库建立索引的方法,包括:在步骤S201,根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标(index indicator),其中所述索引指标用于为所述多个文档建立索引;在步骤S202,将所述多个文档划分为多个文档集合;在步骤S203,根据所述索引指标对所述多个文档集合建立索引。

[0026] 在步骤S201,根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标(index indicator),其中所述索引指标用于为所述多个文档建立索引,具体地,包括:步骤S301,记录用户一段时间内对所述多个文档的各个字段的操作,其中所述操作包括以下操作中的至少一个:查询操作、写操作和分组/排序操作;在步骤S302,统计出所述文档型数据库中的多个文档的各个字段的属性并统计出针对每个属性的相同操作的次数,其中响应于所述各个字段中的字段b的属性包括子属性,针对每个属性的相同操作的次数是针对字段b的属性的全部子属性的相同操作的次数总和;在步骤S303,根据每个属性的相同操作的次数从所述属性中确定索引指标。

[0027] 根据本发明的实施例,示出根据对所述文档型数据库中的3个文档的访问情况确定索引指标,表1、2和3示出所述文档型数据库中的3个文档。图3示出根据本发明实施例的文档1、2和3的属性的树状结构,其中属性human包括三个子属性name、age和children,其中属性children进一步包括子属性name和age。表4记录文档1、2和3的属性和操作,表5统计文档1、2和3的属性和针对每个属性的相同类型的操作次数,图4示出在图3的属性的树状结构上标识出针对每个属性的相同类型的操作次数。

[0028] 表1:文档1

[0029]

序列	字段的属性	字段的取值
1	human	--
2	human/name	Bill
3	human/age	28
4	human/children	--
5	human/children/name	Andy
6	human/children/age	3

表2:文档2

[0030]

序列	字段的属性	字段的取值
1	human	--
2	human/name	Jim
3	human/age	25
4	human/children	--
5	human/children/name	Tommy
6	human/children/age	1

[0031] 表3:文档3

[0032]

序列	字段的属性	字段的取值
1	human	--
2	human/name	Kate
3	human/age	26
4	human/children	--
5	human/children/name	Jack
6	human/children/age	2

[0033] 表4:记录文档1、2和3的属性和操作

[0034]

序号	字段的属性	字段的取值	访问类型
1	human/name	Bill	查询操作
2	human/name	Jim	查询操作
3	human/age	28	写操作
4	human/children/name	Tommy	分组、排序操作
5	human/children/age	25	分组、排序操作
6	human/children/name	Andy	分组、排序操作
7	human/age	2	查询操作
8	human/name	Kate	查询操作
9	human/name	Bill	分组、排序操作
10	human/children/name	Jack	查询操作

[0035] 表5:统计文档1、2和3的属性和针对每个属性的相同类型的操作次数

[0036]

字段的属性	查询操作次数	写操作次数	分组/排序操作次数
human	5	1	4
human/children	1	0	3
human/name	3	0	1
human/age	1	1	0
human/children/nam	1	0	2
human/children/age	0	0	1

[0037] 根据本发明的实施例,可以根据Top-N模式确定索引指标,分别对查询操作、写操作、分组/排序操作赋予权重$\alpha_1,\alpha_2,\alpha_3$,对表5中6个属性的查询操作、写操作、分组/排序操作的操作次数进行加权,针对6个属性计算每个属性的三种操作的操作次数总和的加权值$(W_1,W_2,W_3,W_4,W_5,W_6)$

[0038] $W_1=5\alpha_1+\alpha_2+4\alpha_3$ (1)

[0039] $W_2=\alpha_1+3\alpha_3$ (2)

[0040] $W_3=3\alpha_1+\alpha_3$ (3)

[0041] $W_4=\alpha_1+\alpha_2$ (4)

[0042] $W_5=\alpha_1+2\alpha_3$ (5)

[0043] $W_6=\alpha_3$ (6)

[0044] 把$(W_1,W_2,W_3,W_4,W_5,W_6)$按照从大到小的顺序进行排序,选取对应于前N个加权值的N个属性作为索引指标,其中N可以由非关系型数据库的管理员指定。

[0045] 根据本发明又一实施例,可以根据阈值模式确定索引指标,非关系型数据库的管理员根据经验设定一个阈值T,从$(W_1,W_2,W_3,W_4,W_5,W_6)$中选取超过阈值T的加权值,并将与超过阈值T的加权值对应的属性作为索引指标。

[0046] 由于文档型数据库中文档的数量和种类不断处于变化之中,每间隔一段时间根据记录的用户在一段时间内对所述文档型数据库中的文档的各个字段的操作,统计和更新文档的字段的属性,从中筛选出需要建立索引的索引指标,从而实现索引指标的动态更新。

[0047] 在步骤S202,将所述文档型数据库中的多个文档划分为文档集合。现有技术中针对全部文档建立索引,这样建立索引的时间长、占用内存和磁盘空间,影响数据库系统的性能。根据本发明的实施例,可以将多个文档划分为文档集合,每个文档集合是由至少一个文档构成的,针对文档集合建立索引。例如可以随机将多个文档划分为文档集合,也可以按照用户访问时间、访问频率情况将多个文档划分为文档集合,只要保证文档集合的数目大于或者等于索引指标的数目即可。根据本发明的实施例,文档集合的数目为索引指标的数目的整数倍,对文档集合建立索引算法实现容易、建立时间短、占用内存和磁盘空间小。

[0048] 本领域技术人员理解,步骤S201和步骤S202既可以同时执行,也可以先后执行,在实现上没有先后顺序的限制。

[0049] 在步骤S203,根据所述索引指标对所述文档集合中的文档建立索引,具体地,包括:按照所述索引指标的数目选取文档集合;为每个文档集合分配索引指标并按照分配的索引指标建立索引;响应于确定存在已完成建立索引的文档集合,获取分配给所述完成建

立索引的文档集合的索引指标a;响应于确定存在未分配索引指标的文档集合,选取一个未分配索引指标的文档集合并按照索引指标a建立索引;响应于确定不存在未分配索引指标的文档集合,判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引;响应于判断结果为否,选取一个未分配过索引指标a的文档集合并按照索引指标a建立索引。

[0050] 根据本发明的实施例,进一步包括:获取新增索引指标;响应于确定存在未分配索引指标的文档集合,选取一个未分配索引指标的文档集合并按照新增索引指标建立索引;响应于确定不存在未分配索引指标的文档集合,判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引;响应于判断结果为否,为每个文档集合统计分配的索引指标的数目;为索引指标的分配数目最少的文档集合分配新增索引指标并按照新增索引指标建立索引。

[0051] 根据本发明的实施例,进一步包括:获取废弃索引指标;响应于确定存在已按所述废弃索引指标建立索引的文档集合,删除按照所述废弃索引指标建立的索引并且删除所述废弃索引指标。

[0052] 图5示出根据本发明的实施例根据所述索引指标对所述文档集合中的文档建立索引的流程。首先根据上述步骤S201和S202确定了m个索引指标并且将文档型数据库中的文档划分了n个文档集合(n>m),在步骤S501,从n个文档集合中选取m个文档集合并分配m个索引指标;在步骤S502,按照已分配的索引指标为文档集合建立索引;在步骤S503,查询每个文档建立索引的状态;在步骤S504,判断是否存在已完成建立索引的文档集合;响应于查询结果为否,则返回步骤S503继续查询;响应于查询结果为是,在步骤S505,获取为该文档集合分配的索引指标a;在步骤S506,查询n个文档集合中是否存在未分配索引指标的文档集合;响应于查询结果为是,在步骤S507,随机选取一个未分配索引指标的文档集合并分配索引指标,流程返回步骤S502按照分配的索引指标建立索引,响应于查询结果为否,在步骤S508,判断每个文档集合是否都已按m个索引指标完成建立索引;响应于查询结果为是,流程结束;响应于查询结果为否,在步骤S509,随机选取一个未分配过索引指标a的文档集合;在步骤S510,为选取的文档集合分配索引指标a,流程返回步骤S502按照分配的索引指标建立索引。

[0053] 图6示出根据图5的流程根据新增索引指标建立索引的流程,在步骤S601,获取新增索引指标;在步骤S602,查询是否存在未分配索引指标的文档集合,响应于查询结果为是,在步骤S603,选取一个未分配索引指标的文档集合并分配新增索引指标;在步骤S604,为选取的文档集合按照分配的新增索引指标建立索引,流程返回步骤S601;响应于查询结果为否,在步骤S605,判断是否每个文档集合都已按分配的索引指标完成建立索引;响应于查询结果为是,流程结束;响应于查询结果为否,在步骤S606,统计为每个文档集合分配的索引指标的数目;在步骤S607,为分配索引指标的数目最少的文档集合分配新增索引指标;在步骤S607,该文档集合按照分配的新增索引指标建立索引,流程返回步骤S601。

[0054] 由于数据分块的难度较大,关系型数据库的索引建立是针对表中的全部数据。本发明的实施例不是一次性为文档型数据库中的所有文档按照所有索引指标建立索引,而是将文档型数据库中的多个文档划分为文档集合,针对文档集合按照所有索引指标逐步建立索引,这样可以充分利用文档型数据库易于分块的特点来降低索引建立过程中对数据库系

统性能的影响。按照分块逐步建立索引的方式，在建立过程中随着索引指标动态发生改变例如产生新增索引指标和废弃索引指标，按照废弃索引指标建立的索引被视为无效索引，无效索引的建立可视为性能损耗，因此可有效避免全部文档建立索引产生的性能损耗。

[0055] 本领域技术人员理解，上述步骤S502、S604和S608针对划分的文档集合按照分配的索引指标建立索引的实现方式可以有多种，建立索引的方式不局限于任何特定的索引方式。例如MySQL数据库中常用的四种索引方式：B-Tree索引，Hash索引、Fulltext索引和R-Tree索引，其中B-Tree索引是MySQL数据库使用最为频繁的索引类型，B-树是一种平衡的多路查找树，它在文件系统中很有用。下面给出B-树定义。

[0056] B-树定义：一棵m阶的B-树，或者为空树，或为满足下列特性的m叉树：

[0057] (1)树中每个结点至多有m棵子树；

[0058] (2)若根结点不是叶子结点，则至少有两棵子树；

[0059] (3)根结点之外的所有非终端结点至少有[m/2]棵子树；

[0060] (4)所有的非终端结点中包含以下信息数据：$(n,A_0,K_1,A_1,K_2,\cdots,K_n,A_n)$其中：$K_i(i=1,2,\cdots,n)$为关键码，且$K_i<K_{i+1}$，$A_i$为指向子树根结点的指针$(i=0,1,\cdots,n)$，且指针$A_{i-1}$所指子树中所有结点的关键码均小于$K_i(i=1,2,\cdots,n)$，$A_n$所指子树中所有结点的关键码均大于$K_n$，$n([m/2])-1\leq n\leq m-1)$为关键码的个数。

[0061] (5)所有的叶子结点都出现在同一层次上，并且不带信息(可以看作是外部结点或查找失败的结点，实际上这些结点不存在，指向这些结点的指针为空)。

[0062] B+树是应文件系统所需而产生的一种B-树的变形树。一棵m阶的B+树和m阶的B-树的差异在于：

[0063] (1)有n棵子树的结点中含有n个关键码；

[0064] (2)所有的叶子结点中包含了全部关键码的信息，及指向含有这些关键码记录的指针，且叶子结点本身依关键码的大小自小而大的顺序链接。

[0065] (3)所有的非终端结点可以看成是索引部分，结点中仅含有其子树根结点中最大(或最小)关键码。

[0066] 尽管在上文中介绍了B-树和B+树，但是建立索引的方式并不局限于此，只要能够实现非关系型数据库建立索引，现有的以及未来的可能实现数据库建立索引的方式都在本发明的范围内。

[0067] 图7示出根据图5的流程处理废弃索引指标的流程，在步骤S701，获取废弃索引指标，确定索引指标的过程是动态更新的过程，每间隔一段时间重新确定索引指标，如果重新确定的索引指标中不包括之前确定的索引指标，这种情况，之前确定的索引指标就成为废弃索引指标；在步骤S702，判断是否存在已按废弃索引指标建立索引的文档集合；响应于判断结果为是，其中判断结果为是的文档集合包括两种情形，一种是已按废弃索引指标完成建立索引的文档集合，一种是正在按照废弃索引指标建立索引的文档集合。在步骤S703，删除按照废弃索引指标建立的索引，其中包括按照废弃索引指标全部完成建立的索引，以及按照废弃索引指标部分完成建立的索引；在步骤S704，删除废弃索引指标；响应于判断结果为否，前进到步骤S704，流程结束。

[0068] 前面已经参考附图描述了实现本发明的方法的各个实施例。本领域技术人员可以理解的是，上述方法可以以软件方式实现，也可以以硬件方式实现，或者通过软件与硬件相

结合的方式实现。并且，本领域技术人员可以理解，通过以软件、硬件或者软硬件相结合的方式实现上述方法中的各个步骤，可以提供一种文档型数据库建立索引的装置。即使该装置在硬件结构上与通用处理设备相同，由于其中所包含的软件的作用，使得该装置表现出区别于通用处理设备的特性，从而形成本发明的各个实施例的装置。

[0069] 基于同一发明构思，根据本发明的实施例还提出一种文档型数据库建立索引的装置，图8示出根据本发明实施例的文档型数据库建立索引的装置800，该装置包括：第一确定模块801，被配置为根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标，其中所述索引指标用于为所述多个文档建立索引；划分模块802，被配置为将所述文档型数据库中的多个文档划分为多个文档集合；建立模块803，被配置为根据所述索引指标对所述多个文档集合建立索引。

[0070] 根据本发明的实施例，其中第一确定模块进一步包括：记录模块，被配置为记录一段时间内用户对所述多个文档的各个字段的操作；第一统计模块，被配置为统计出所述多个文档的各个字段的属性并统计出针对每个属性的相同操作的次数；第二确定模块，被配置为根据每个属性的相同操作的次数从所述属性中确定索引指标。

[0071] 根据本发明的实施例，其中响应于所述各个字段中的字段b的属性包括子属性，针对每个属性的相同类型的操作次数是针对所述字段b的属性的全部子属性的相同操作的次数总和。

[0072] 根据本发明的实施例，其中用户对所述文档型数据库中的多个文档的各个字段的操作包括以下操作中的至少一个：查询操作、写操作、分组/排序操作。

[0073] 根据本发明的实施例，其中所述多个文档集合的数目大于或者等于索引指标的数目。根据本发明的实施例，其中建立模块803包括：第一选取模块，被配置为按照所述索引指标的数目选取文档集合；第一分配模块，被配置为为选取的每个文档集合分配索引指标并按照分配的索引指标建立索引；第一获取模块，被配置为响应于确定存在已完成建立索引的文档集合，获取分配给所述完成建立索引的文档集合的索引指标a；第二选取模块，被配置为响应于确定存在未分配索引指标的文档集合，选取一个未分配索引指标的文档集合并按照索引指标a建立索引。

[0074] 根据本发明的实施例，，其中建立模块803进一步包括：第一判断模块，被配置为响应于确定不存在未分配索引指标的文档集合，判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引；第三选取模块，被配置为响应于判断结果为否，选取一个未分配过索引指标a的文档集合并按照索引指标a建立索引。

[0075] 根据本发明的实施例，该装置进一步包括：第二获取模块，被配置为获取新增索引指标；第四选取模块，被配置为响应于确定存在未分配索引指标的文档集合，选取一个未分配索引指标的文档集合并按照新增索引指标建立索引；

[0076] 根据本发明的实施例，该装置进一步包括：第二判断模块，被配置为响应于确定不存在未分配索引指标的文档集合，判断所述多个文档集合的每一个是否都已按全部所述索引指标完成建立索引；第二统计模块，被配置为响应于判断结果为否，为每个文档集合统计分配的索引指标的数目；第二分配模块，被配置为为索引指标的分配数目最少的文档集合分配新增索引指标并按照新增索引指标建立索引。

[0077] 根据本发明的实施例，该装置进一步包括：第三获取模块，被配置为获取废弃索引

指标;删除模块,被配置为响应于确定存在已按所述废弃索引指标建立索引的文档集合,删除按照所述废弃索引指标建立的索引并且删除所述废弃索引指标。

[0078] 上述每个模块的具体实现方法参照根据本发明实施例的文档型数据库建立索引的方法的详细描述,在此不一一赘述。

[0079] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0080] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0081] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0082] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0083] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0084] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据

处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0085] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0086] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0087] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

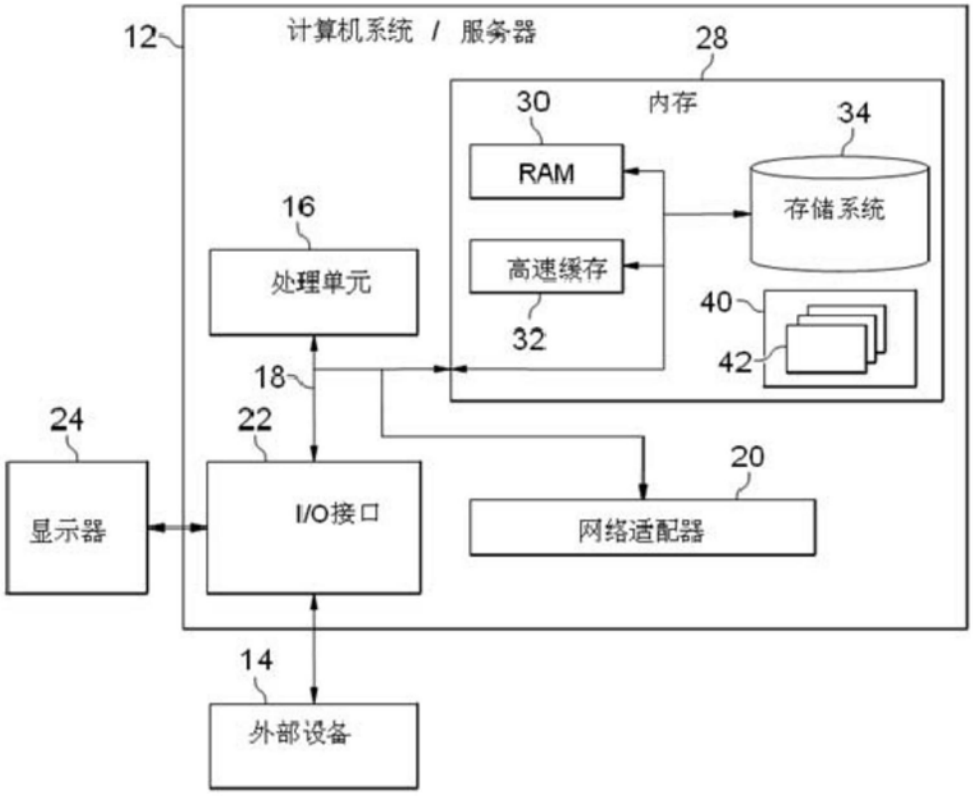

12
计算机系统 / 服务器
28
30
内存
34
RAM
存储系统
16
处理单元
高速缓存
32
40
42
18
24
22
20
显示器
I/O接口
网络适配器
14
外部设备

图1

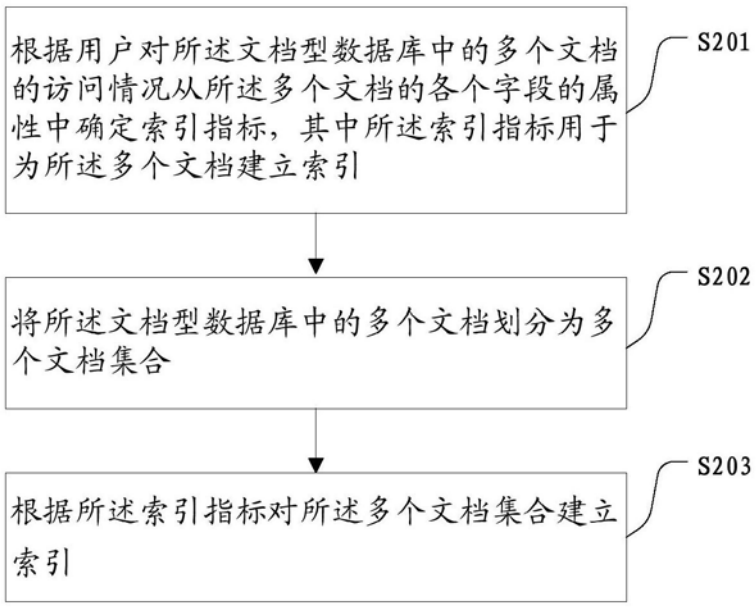

根据用户对所述文档型数据库中的多个文档的访问情况从所述多个文档的各个字段的属性中确定索引指标，其中所述索引指标用于为所述多个文档建立索引
S201
将所述文档型数据库中的多个文档划分为多个文档集合
S202
根据所述索引指标对所述多个文档集合建立索引
S203

图2

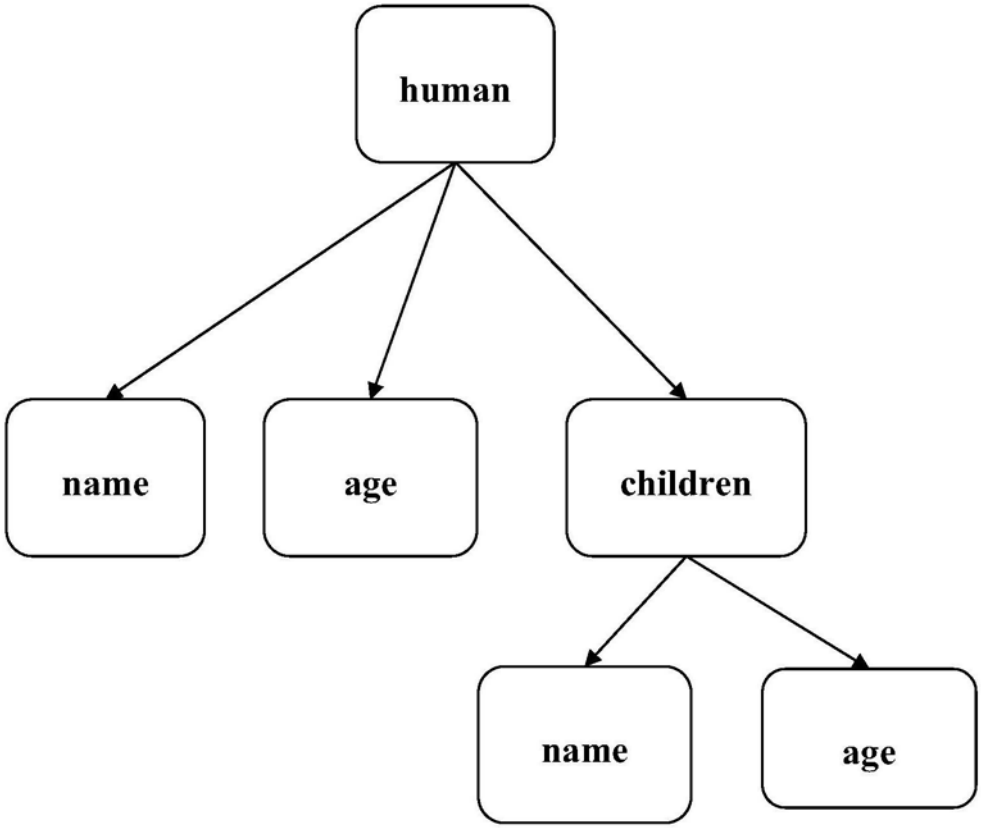


图3

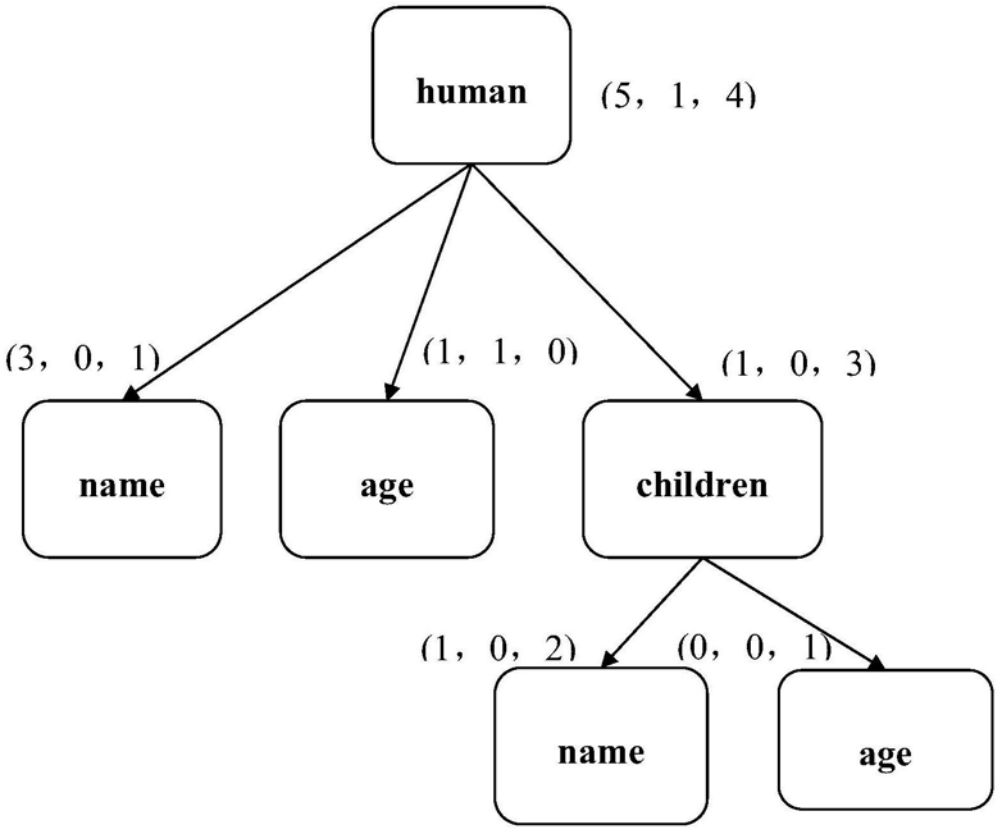


图4

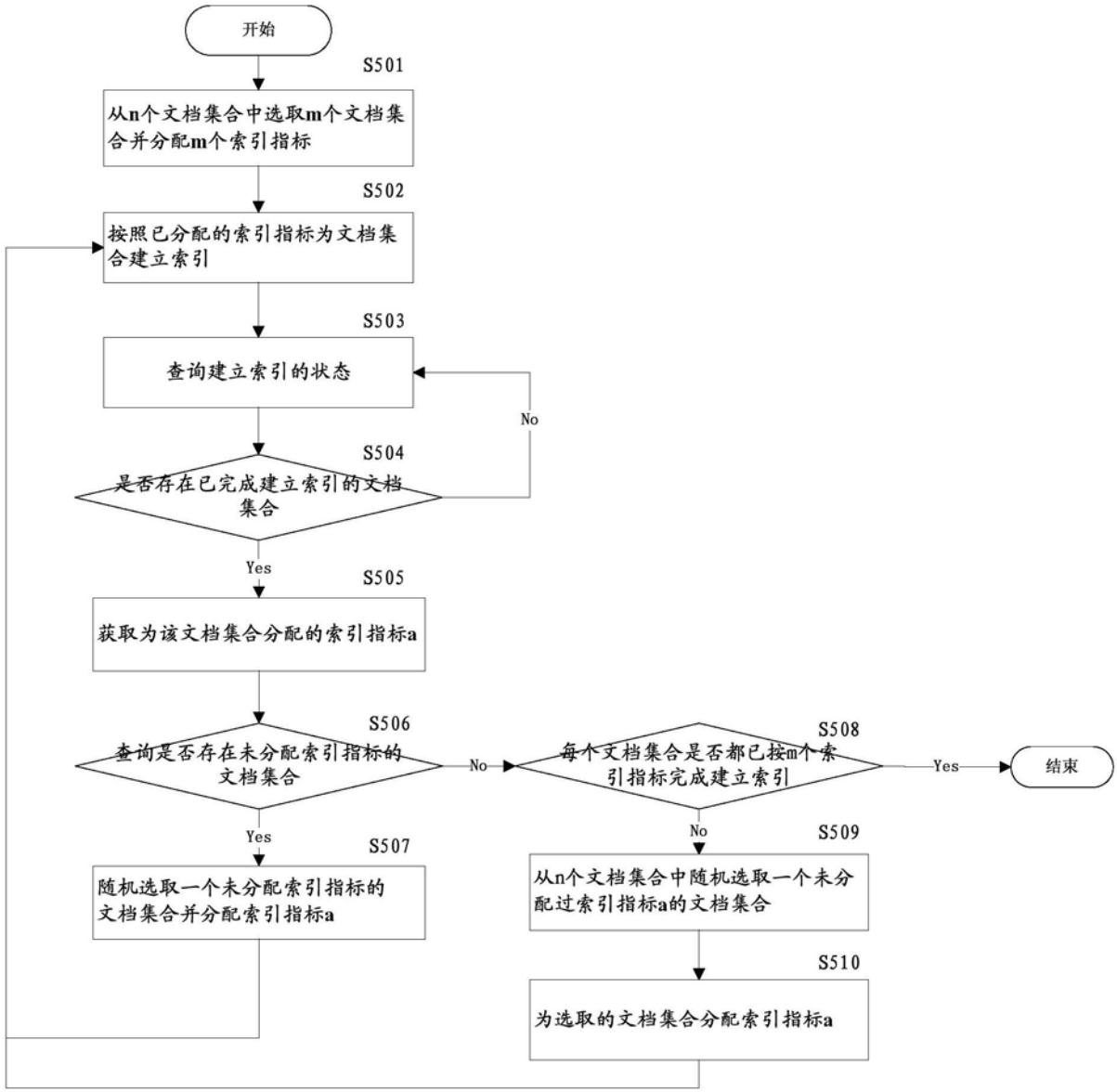

开始
S501
从n个文档集合中选取m个文档集合并分配m个索引指标
S502
按照已分配的索引指标为文档集合建立索引
S503
查询建立索引的状态
No
S504
是否存在已完成建立索引的文档集合
Yes
S505
获取为该文档集合分配的索引指标a
S506
查询是否存在未分配索引指标的文档集合
No
S508
每个文档集合是否都已按m个索引指标完成建立索引
Yes
结束
Yes
S507
随机选取一个未分配索引指标的文档集合并分配索引指标a
No
S509
从n个文档集合中随机选取一个未分配过索引指标a的文档集合
S510
为选取的文档集合分配索引指标a

图5

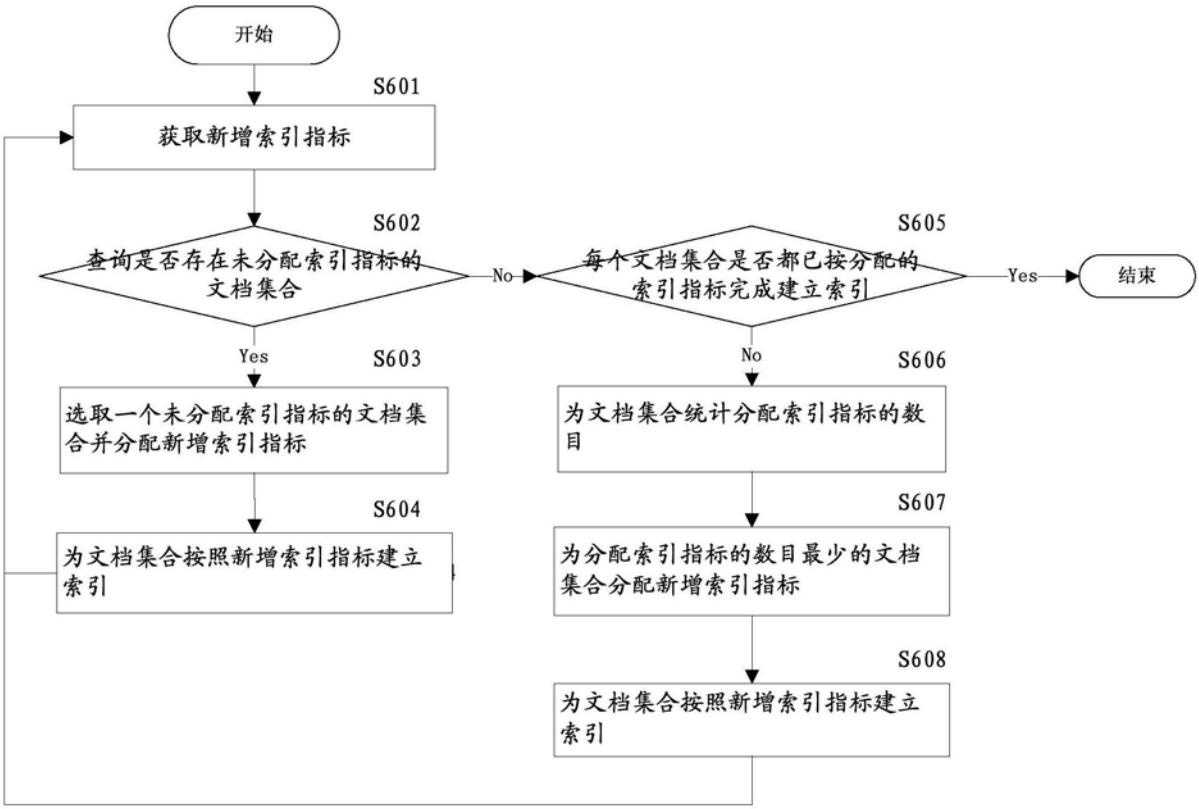

开始
S601
获取新增索引指标
S602
查询是否存在未分配索引指标的文档集合
No
S605
每个文档集合是否都已按分配的索引指标完成建立索引
Yes
结束
Yes
S603
选取一个未分配索引指标的文档集合并分配新增索引指标
S604
为文档集合按照新增索引指标建立索引
No
S606
为文档集合统计分配索引指标的数目
S607
为分配索引指标的数目最少的文档集合分配新增索引指标
S608
为文档集合按照新增索引指标建立索引

图6

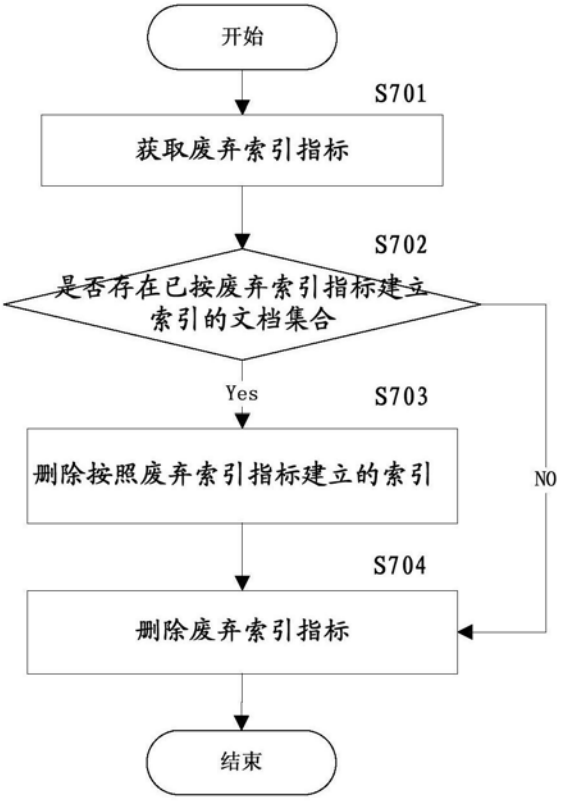

开始
S701
获取废弃索引指标
S702
是否存在已按废弃索引指标建立索引的文档集合
Yes
S703
删除按照废弃索引指标建立的索引
NO
S704
删除废弃索引指标
结束

图7

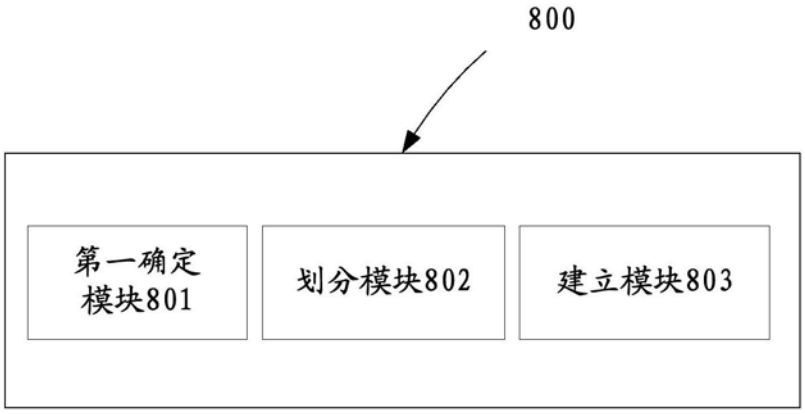

800
第一确定模块801
划分模块802
建立模块803

图8