



(12) 发明专利申请

(10) 申请公布号 CN 103198119 A

(43) 申请公布日 2013. 07. 10

(21) 申请号 201310112125. 9

(22) 申请日 2013. 04. 02

(71) 申请人 浪潮电子信息产业股份有限公司

地址 250014 山东省济南市高新区舜雅路  
1036 号

(72) 发明人 王通 郭鹏

(51) Int. Cl.

G06F 17/30(2006. 01)

G06F 9/38(2006. 01)

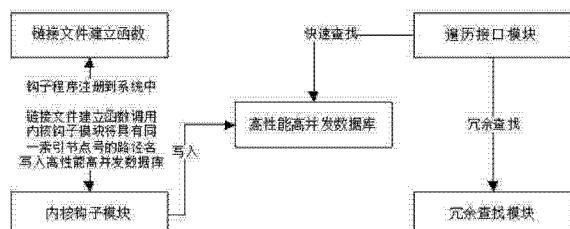
权利要求书1页 说明书2页 附图2页

(54) 发明名称

一种快速查找具有相同重复数据删除标识的所有链接文件的方法

(57) 摘要

本发明提供一种快速查找具有相同重复数据删除标识的所有链接文件的方法，是以高性能高并发数据库为核心，通过整合遍历接口、内核钩子模块和冗余查找模块，使该查找方法达到比较高的效率，该方法的模块结构包括：高性能高并发数据库(1)，内核钩子模块(2)、遍历接口模块(3)、冗余查找模块(4)。内核钩子模块、遍历接口模块、冗余查找模块支持高并发的多进程多线程操作，从而提高系统的整体性能。冗余查找模块提供了冗余配置，从而提高系统的高可用性。很少需要遍历整个文件系统目录树进行查找，极为高效。



1. 一种快速查找具有相同重复数据删除标识的所有链接文件的方法，其特征在于，以高性能高并发数据库为核心，通过整合遍历接口、内核钩子模块和冗余查找模块，使该查找方法达到比较高的效率，该方法的模块结构包括：高性能高并发数据库(1)，内核钩子模块(2)、遍历接口模块(3)、冗余查找模块(4) 其中：

高性能高并发数据库(1)是结构的核心，负责存放大量的链接文件路径信息，并支持多进程、多线程高并发访问；

内核钩子模块(2)主要负责建立链接文件时的信息收集及信息存放，支持多线程并发；

遍历接口模块(3)为上层应用程序遍历系统提供调用接口；

冗余查找模块(4)的作用为在高性能高并发数据库(1)中没有所需要的信息时，遍历整个存储系统，进行冗余查找，并将查找到的信息放入高性能高并发数据库(1)中。

2. 根据权利要求 1 所述的方法，其特征在于内核钩子模块、遍历接口模块、冗余查找模块支持高并发的多进程多线程操作，从而提高系统的整体性能。

3. 根据权利要求 1 所述的方法，其特征在于冗余查找模块提供了方法的冗余配置，从而提高系统的高可用性。

# 一种快速查找具有相同重复数据删除标识的所有链接文件的方法

## 技术领域

[0001] 本发明涉及计算机应用技术领域,具体涉及一种快速查找具有相同重复数据删除标识的所有链接文件的方法。

## 背景技术

[0002] 进入 21 世纪以来,随着信息时代的加速,企业数据呈现出爆炸性增长的趋势,特别是移动互联网、物联网和云计算的发展更加剧了数据的爆炸式增长。IDC 报告指出,全球数据量每年以 60% 的速度递增,2010 年全球数据量达 1.8ZB,2015 年将达到 8ZB,2020 年将达到 35ZB,标志着“大数据”时代的到来。数据增长带来如下巨大的问题:成本急剧增加、带宽压力大、耗能问题严重、设备空间占用巨大、靠增加设备无法彻底解决数据量激增的问题等问题,同时,世界所面临的能源问题日益严峻,在高科技的 IT 领域能源浪费和环保更加引人注目。互联网的广泛使用让大型企业、政府机关、金融机构的信息中心规模日益膨胀,数据交换增加,设备堆积成山,占地面积越来越多,耗电量屡创新高。为实现信息和管理优化,在构建企业信息架构时,更加呼吁绿色的节能技术。节约能源,减少电力消耗,降低系统成本,急需研究面向新兴应用的新型绿色存储技术。在这个大趋势下,重复数据删除技术蕴育而生,重复数据删除技术能够有效地减少用户存储系统中的重复数据,从而为用户节省了存储容量,降低存储成本和管理难度。

[0003] 现有的查找具有同一重复数据删除标识的所有链接文件方法都必须逐次遍历整个文件系统目录树,并对每一个查找到的文件,获取其标识并进行比较,对于十亿级别文件目录的遍历将耗费大量的时间和资源,在数据重删技术中,按照重删的方法可以分为:文件级重删和块级重删。在文件级的重删方案中,需要对内容重复的文件保存一个副本,并在重复文件所在的路径处建立到这个副本的链接(包含证明文件内容一致的重复数据删除标识,一般是文件内容的哈希值)。当需要快速恢复具有同一文件内容的多个路径下的文件时,如何快速查找到具有相同内容的所有文件链接路径的方法就极为重要。

## 发明内容

[0004] 本发明的目的是提供一种快速查找具有相同重复数据删除标识的所有链接文件的方法。

[0005] 现有的查找具有同一重复数据删除标识的所有链接文件方法都必须逐次遍历整个文件系统目录树,并对每一个查找到的文件,获取其标识并进行比较,对于十亿级别文件目录的遍历将耗费大量的时间和资源。

[0006] 本发明的目的是按以下方式实现的:

本发明的结构是高性能高并发数据库为中心的方法,该系统体系结构包括:高性能高并发数据库(1),内核钩子模块(2)、遍历接口模块(3)、冗余查找模块(4),内核钩子模块、遍历接口模块、冗余查找模块支持高并发的多进程多线程操作,从而提高系统的整体性能,

其中：

高性能高并发数据库(1)是体系结构的核心,负责存放大量的硬链接信息,并支持多进程、多线程高并发访问；

内核钩子模块(2)主要负责建立链接文件时的信息收集及信息存放,支持多线程并发；

遍历接口模块(3)为上层应用程序遍历系统提供调用接口；

冗余查找模块(4)的作用为在高性能高并发数据库(1)中没有所需要的信息时,遍历整个存储系统,进行冗余查找,并将查找到的信息放入高性能高并发数据库(1)中。

[0007] 本发明的有益效果是：内核钩子模块、遍历接口模块、冗余查找模块支持高并发的多进程多线程操作,从而提高系统的整体性能。冗余查找模块提供了冗余配置,从而提高系统的高可用性。很少需要遍历整个文件系统目录树进行查找,极为高效。

## 附图说明

[0008] 图1是传统的查找具有同一标识的所有硬链接路径拓扑图；

图2是快速查找具有相同重复数据文件标识的所有链接文件流程示意图。

## 具体实施方式

[0009] 参照说明书附图对本发明的方法作以下详细地说明。

[0010] 正如发明内容中所描述的,本发明体系结构主要包括:高性能高并发数据库(1),内核钩子模块(2)、遍历接口模块(3)、冗余查找模块(4)。

[0011] 我们提出的基于高性能高并发数据库的快速查找具有一种快速查找具有相同重复数据删除标识的所有链接文件方法以高性能高并发数据库为核心,其特征在于在方法中,内核钩子模块、遍历接口模块、冗余查找模块支持高并发的多进程多线程操作,从而提高系统的整体性能。内核钩子模块、遍历接口模块、冗余查找模块进行冗余配置,从而提高系统的高可用性。如图2所示,本系统体系结构主要包括:高性能高并发数据库(1),内核钩子模块(2)、遍历接口模块(3)、冗余查找模块(4)。

[0012] 高性能高并发数据库作为此方法的核心,起到信息存储和高速并发查找等作用。

[0013] 内核钩子模块注册进内核,建立链接文件的函数转入内核执行时,使用内核钩子程序,将文件路径及重复数据删除标识等信息存放入高性能高并发数据库,并将重复数据删除标识写入链接文件。

[0014] 遍历接口模块提供遍历调用的接口,是各种查找函数的入口,查找时,首先进入高性能高并发数据库进行查找,如果能够找到数据库键值与查找的标识匹配,则将该键值对应的内容返回给调用函数,否则进入冗余查找模块进行查找。

[0015] 冗余查找模块将以深度遍历或广度遍历方法遍历整个文件系统目录树,对每个文件获取其标识并和查找关键字对比,直到遍历完整个文件系统目录树,将得到的结果返回。

[0016] 除说明书所述的技术特征外,均为本专业技术人员的已知技术。

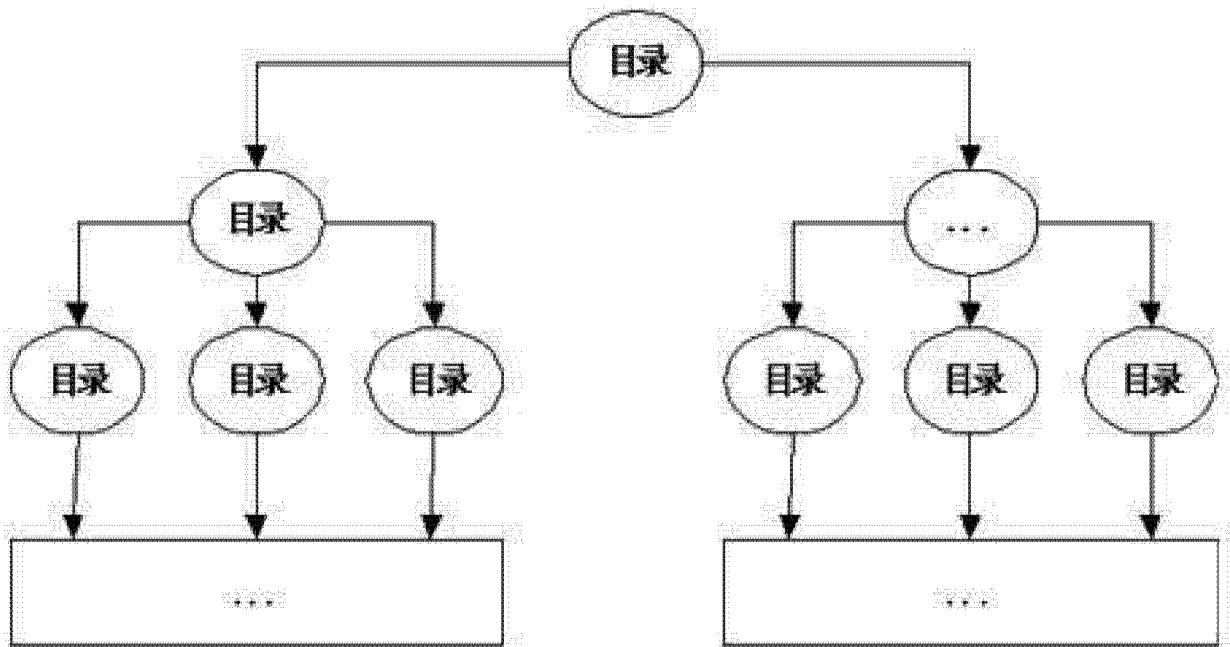


图 1

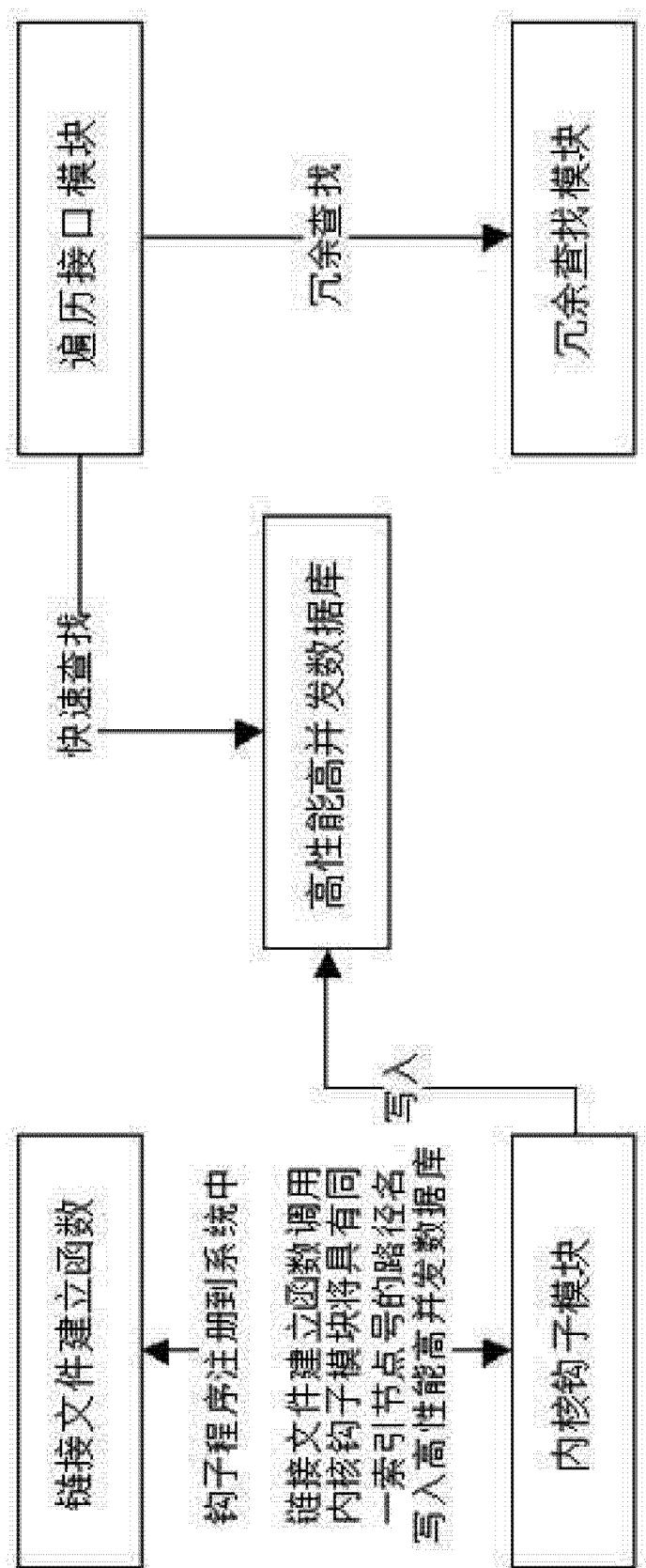


图 2