(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 107704472 B
(45)授权公告日 2020.07.24

(21)申请号 201610648299.0

(22)申请日 2016.08.09

(65)同一申请的已公布的文献号
申请公布号 CN 107704472 A

(43)申请公布日 2018.02.16

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 关坤 冷继南 沈建强 王工艺

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 申健

(51)Int.Cl.
G06F 16/13(2019.01)
G06F 16/14(2019.01)
G06F 16/174(2019.01)

(56)对比文件
CN 102156727 A,2011.08.17,
CN 101963982 A,2011.02.02,
CN 104408154 A,2015.03.11,
CN 103514250 A,2014.01.15,
US 2015039572 A1,2015.02.05,

审查员 沈婷婷

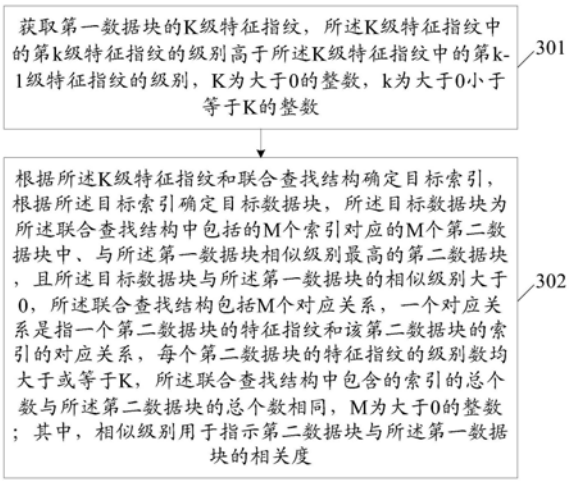
权利要求书4页 说明书13页 附图5页

(54)发明名称

一种查找数据块的方法及装置

(57)摘要

本发明实施例公开了一种查找数据块的方法及装置，涉及数据检测领域，用以解决现有技术中的多级查找结构中存在大量的数据冗余的问题。该方法包括：获取第一数据块的K级特征指纹；根据K级特征指纹和联合查找结构确定目标索引，根据目标索引确定目标数据块，目标数据块为联合查找结构中包括的M个索引对应的M个第二数据块中、与第一数据块相似级别最高的第二数据块，且目标数据块与第一数据块的相似级别大于0，联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，联合查找结构中包含的索引的总个数与第二数据块的总个数相同。



CN 107704472 B

1.一种查找数据块的方法,其特征在于,包括:

获取第一数据块的K级特征指纹,所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别,K为大于0的整数,k为大于0小于等于K的整数;

根据所述K级特征指纹和联合查找结构确定目标索引,根据所述目标索引确定目标数据块,所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块,且所述目标数据块与所述第一数据块的相似级别大于0,所述联合查找结构包括M个对应关系,一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系,每个第二数据块的特征指纹的级别数均大于或等于K,所述联合查找结构中包含的索引的总个数与第二数据块的总个数相同,M为大于0的整数;其中,相似级别用于指示第二数据块与所述第一数据块的相关度。

2.根据权利要求1所述的方法,其特征在于,所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹,或者,所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

3.根据权利要求2所述的方法,其特征在于,所述根据所述K级特征指纹和联合查找结构确定目标索引,包括:

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配;

若匹配成功,确定所述第一对应关系中的索引为目标索引;

若匹配失败、且所述K级特征指纹大于所述第一对应关系中的特征指纹,记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别,确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹大的对应关系为新的第一对应关系,将所述K级特征指纹与新的第一对应关系的特征指纹继续进行匹配;若匹配失败、且所述K级特征指纹小于所述第一对应关系中的特征指纹,记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别,确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹小的对应关系为新的第一对应关系,将所述K级特征指纹与新的第一对应关系继续进行匹配,直至确定出目标索引,若全部的第一对应关系均与所述K级特征指纹匹配失败,所述目标索引为所述全部的第一对应关系中的与所述K级特征指纹匹配级别最高的第一对应关系中的索引,初始的第一对应关系为所述联合查找结构中的第1个对应关系;

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配,包括:从所述K级特征指纹中的第1级特征指纹开始,依次与所述第一对应关系中的相同级别的特征指纹进行匹配,若所述K级特征指纹中的第K级特征指纹与所述第一对应关系中的第K级特征指纹匹配成功,则匹配成功,否则,匹配失败。

4.根据权利要求2所述的方法,其特征在于,所述根据所述K级特征指纹和联合查找结构确定目标索引,包括:

将所述K级特征指纹中的第a级特征指纹与所述联合查找结构中的M个对应关系中的第a级特征指纹进行匹配,确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系,令a＝a+1,将所述K级特征指纹中的第a级特征指纹与所述全部对应关系中的第a级特征指纹进行匹配,确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系,直至确定

出与所述K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

5.根据权利要求1-4任一项所述的方法，其特征在于，所述方法还包括：

若所述目标数据块为与所述第一数据块相同的数据块，存储指示信息，所述指示信息用于指示所述目标数据块的地址；

若所述目标数据块为与所述第一数据块相似的数据块，基于所述目标数据块对所述第一数据块进行相似压缩，并存储压缩后的数据；

若未确定出所述目标数据块，存储所述第一数据块。

6.根据权利要求1-4任一项所述的方法，其特征在于，所述方法还包括：

计算所述第一数据块的特征指纹，将所述第一数据块的特征指纹与所述第一数据块的索引的对应关系添加至所述联合查找结构中。

7.一种查找数据块的装置，其特征在于，包括：

获取单元，用于获取第一数据块的K级特征指纹，所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数；

确定单元，用于根据所述K级特征指纹和联合查找结构确定目标索引，根据所述目标索引确定目标数据块，所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块，且所述目标数据块与所述第一数据块的相似级别大于0，所述联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，所述联合查找结构中包含的索引的总个数与第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与所述第一数据块的相关度。

8.根据权利要求7所述的装置，其特征在于，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹，或者，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

9.根据权利要求8所述的装置，其特征在于，所述确定单元，具体用于：

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配；

若匹配成功，确定所述第一对应关系中的索引为目标索引；

若匹配失败、且所述K级特征指纹大于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹大的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且所述K级特征指纹小于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹小的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与所述K级特征指纹匹配失败，所述目标索引为所述全部的第一对应关系中的与所述K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为所述联合查找结构中的第1个对应关系；

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配，

包括:从所述K级特征指纹中的第1级特征指纹开始,依次与所述第一对应关系中的相同级别的特征指纹进行匹配,若所述K级特征指纹中的第K级特征指纹与所述第一对应关系中的第K级特征指纹匹配成功,则匹配成功,否则,匹配失败。

10.根据权利要求8所述的装置,其特征在于,所述确定单元,具体用于:

将所述K级特征指纹中的第a级特征指纹与所述联合查找结构中的M个对应关系中的第a级特征指纹进行匹配,确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系,令a=a+1,将所述K级特征指纹中的第a级特征指纹与所述全部对应关系中的第a级特征指纹进行匹配,确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系,直至确定出与所述K级特征指纹匹配级别最高的对应关系中的索引为目标索引,a的初始值为1。

11.根据权利要求7-10任一项所述的装置,其特征在于,所述装置还包括压缩存储单元;

所述压缩存储单元,用于当所述目标数据块为与所述第一数据块相同的数据块时,存储指示信息,所述指示信息用于指示所述目标数据块的地址;或者,用于当所述目标数据块为与所述第一数据块相似的数据块时,基于所述目标数据块对所述第一数据块进行相似压缩,并存储压缩后的数据;或者,用于当未确定出所述目标数据块时,存储所述第一数据块。

12.根据权利要求7-10任一项所述的装置,其特征在于,所述装置还包括执行单元;

所述执行单元,用于计算所述第一数据块的特征指纹,将所述第一数据块的特征指纹与所述第一数据块的索引的对应关系添加至所述联合查找结构中。

13.一种查找数据块的装置,其特征在于,包括:存储器和处理器,所述存储器用于存储一组代码,所述处理器用于根据该组代码执行以下动作:

获取第一数据块的K级特征指纹,所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别,K为大于0的整数,k为大于0小于等于K的整数;

根据所述K级特征指纹和联合查找结构确定目标索引,根据所述目标索引确定目标数据块,所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块,且所述目标数据块与所述第一数据块的相似级别大于0,所述联合查找结构包括M个对应关系,一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系,每个第二数据块的特征指纹的级别数均大于或等于K,所述联合查找结构中包含的索引的总个数与第二数据块的总个数相同,M为大于0的整数;其中,相似级别用于指示第二数据块与所述第一数据块的相关度。

14.根据权利要求13所述的装置,其特征在于,所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹,或者,所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

15.根据权利要求14所述的装置,其特征在于,所述处理器,具体用于:

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配;

若匹配成功,确定所述第一对应关系中的索引为目标索引;

若匹配失败、且所述K级特征指纹大于所述第一对应关系中的特征指纹,记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别,确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹大的对应关系为新的第一对应

关系，将所述K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且所述K级特征指纹小于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹小的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与所述K级特征指纹匹配失败，所述目标索引为所述全部的第一对应关系中的与所述K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为所述联合查找结构中的第1个对应关系；

将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从所述K级特征指纹中的第1级特征指纹开始，依次与所述第一对应关系中的相同级别的特征指纹进行匹配，若所述K级特征指纹中的第K级特征指纹与所述第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

16.根据权利要求14所述的装置，其特征在于，所述处理器，具体用于：

将所述K级特征指纹中的第a级特征指纹与所述联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将所述K级特征指纹中的第a级特征指纹与所述全部对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与所述K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

17.根据权利要求13-16任一项所述的装置，其特征在于，所述处理器，还用于：

当所述目标数据块为与所述第一数据块相同的数据块时，存储指示信息，所述指示信息用于指示所述目标数据块的地址；

当所述目标数据块为与所述第一数据块相似的数据块时，基于所述目标数据块对所述第一数据块进行相似压缩，并存储压缩后的数据；

当未确定出所述目标数据块时，存储所述第一数据块。

18.根据权利要求13-16任一项所述的装置，其特征在于，所述处理器，还用于：

计算所述第一数据块的特征指纹，将所述第一数据块的特征指纹与所述第一数据块的索引的对应关系添加至所述联合查找结构中。

一种查找数据块的方法及装置

技术领域

[0001] 本发明实施例涉及数据检测领域，尤其涉及一种查找数据块的方法及装置。

背景技术

[0002] 数据检测技术广泛应用于互联网、图像识别、大数据分析和数据缩减等技术领域，其中，相同和/或相似数据查找是数据检测技术中的重要环节。目前，基于单一特征指纹的数据查找可以采用查找树、哈希表查找等成熟的查找方法进行查找，由于根据单一特征指纹查找相同数据或相似数据，必然无法提升数据的压缩率，并且在数据缩减领域，经常会出现基于多个特征指纹的数据查找场景，例如，重删&delta压缩技术场景，因此，需要部署多级查找结构对相同数据和相似数据进行查找。

[0003] 目前的多级查找结构如图1所示，图1中所示的多级查找结构为N(N为大于1的整数)级查找结构，一级查找结构为一个查找结构，每个查找结构中包括M(M为大于1的整数)个特征指纹和与M个特征指纹一一对应的M个指针，这M个指针为与M个数据块一一对应的M个指针，数据块的指针用于指向该数据块的地址，如图1所示，Fpn-Bm表示M个数据块中的第m(m为大于0小于等于M的整数)个数据块的第n(n为大于0小于等于N的整数)个特征指纹，I-Bm表示第m个数据块的指针，Bm表示第m个数据块，参见图1可知，由于一个数据块的N个特征指纹各对应一个该数据块的指针，因此，每个查找结构中都需要包括M个指针，存在着大量的冗余数据。

发明内容

[0004] 本发明的实施例提供了一种查找数据块的方法及装置，用以解决现有技术中的多级查找结构中存在大量的数据冗余的问题。

[0005] 为达到上述目的，本发明的实施例采用如下技术方案：

[0006] 第一方面，提供了一种查找数据块的方法，包括：获取第一数据块的K级特征指纹，K级特征指纹中的第k级特征指纹的级别高于K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数；根据K级特征指纹和联合查找结构确定目标索引，根据目标索引确定目标数据块，目标数据块为联合查找结构中包括的M个索引对应的M个第二数据块中、与第一数据块相似级别最高的第二数据块，且目标数据块与第一数据块的相似级别大于0，联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，联合查找结构中包含的索引的总个数与第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与第一数据块的相关度。

[0007] 第一方面提供的方法可以通过联合查找结构对相同数据或相似数据进行查找，由于联合查找结构中包含的索引的总个数与第二数据块的总个数相同，一个第二数据块的全部的特征指纹对应唯一一个该第二数据块的索引，因此，与现有技术中的方案相比，可以大大的减少数据的冗余。

[0008] 结合第一方面，在第一种可能的实现方式中，第一数据块和第二数据块的特征指纹包括相似特征指纹，或者，第一数据块和第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

[0009] 结合第一方面的第一种可能的实现方式，在第二种可能的实现方式中，根据K级特征指纹和联合查找结构确定目标索引，包括：将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配；若匹配成功，确定第一对应关系中的索引为目标索引；若匹配失败、且K级特征指纹大于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹大的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且K级特征指纹小于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹小的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与K级特征指纹匹配失败，目标索引为全部的第一对应关系中的与K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为联合查找结构中的第1个对应关系；将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从K级特征指纹中的第1级特征指纹开始，依次与第一对应关系中的相同级别的特征指纹进行匹配，若K级特征指纹中的第K级特征指纹与第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

[0010] 结合第一方面的第一种可能的实现方式，在第三种可能的实现方式中，根据K级特征指纹和联合查找结构确定目标索引，包括：将K级特征指纹中的第a级特征指纹与联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将K级特征指纹中的第a级特征指纹与全部对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

[0011] 采用上述第二种可能的实现方式和第三种可能的实现方式进行数据块的查找，不需要将第一数据块的K级特征指纹与全部的第二数据块的特征指纹进行匹配，能够提高查找效率。

[0012] 结合第一方面、第一方面的第一种可能的实现方式至第三种可能的实现方式中的任一种，在第四种可能的实现方式中，该方法还包括：若目标数据块为与第一数据块相同的数据块，存储指示信息，指示信息用于指示目标数据块的地址；若目标数据块为与第一数据块相似的数据块，基于目标数据块对第一数据块进行相似压缩，并存储压缩后的数据；若未确定出目标数据块，存储第一数据块。

[0013] 该第四种可能的实现方式，在确定目标数据块之后，若目标数据块为与第一数据块相同的数据块，可以通过存储指示信息的方式实现数据去重，数据去重是一种数据无损的冗余数据缩减技术，使得多个相同的数据块在存储系统中只存储一个数据块副本，减少了存储数据所需的资源，节约了成本；若目标数据块为与第一数据块相似的数据块，在确定目标数据块之后，基于目标数据块对第一数据块进行相似压缩，可以减少存储的数据量，提

高数据的压缩率，节约存储空间。

[0014] 结合第一方面、第一方面的第一种可能的实现方式至第四种可能的实现方式中的任一种，在第五种可能的实现方式中，该方法还包括：计算第一数据块的特征指纹，将第一数据块的特征指纹与第一数据块的索引的对应关系添加至联合查找结构中。

[0015] 第二方面，提供了一种查找数据块的装置，包括：获取单元，用于获取第一数据块的K级特征指纹，K级特征指纹中的第k级特征指纹的级别高于K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数；确定单元，用于根据K级特征指纹和联合查找结构确定目标索引，根据目标索引确定目标数据块，目标数据块为联合查找结构中包括的M个索引对应的M个第二数据块中、与第一数据块相似级别最高的第二数据块，且目标数据块与第一数据块的相似级别大于0，联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，联合查找结构中包含的索引的总个数与第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与第一数据块的相关度。

[0016] 第二方面提供的装置中的各个单元用于执行第一方面提供的方法，因此，该装置的有益效果可以参见上述方法部分的有益效果，在此不再赘述。

[0017] 结合第二方面，在第一种可能的实现方式中，第一数据块和第二数据块的特征指纹包括相似特征指纹，或者，第一数据块和第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

[0018] 结合第二方面的第一种可能的实现方式，在第二种可能的实现方式中，确定单元，具体用于：将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配；若匹配成功，确定第一对应关系中的索引为目标索引；若匹配失败、且K级特征指纹大于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹大的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且K级特征指纹小于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹小的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与K级特征指纹匹配失败，目标索引为全部的第一对应关系中的与K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为联合查找结构中的第1个对应关系；将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从K级特征指纹中的第1级特征指纹开始，依次与第一对应关系中的相同级别的特征指纹进行匹配，若K级特征指纹中的第K级特征指纹与第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

[0019] 结合第二方面的第一种可能的实现方式，在第三种可能的实现方式中，确定单元，具体用于：将K级特征指纹中的第a级特征指纹与联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将K级特征指纹中的第a级特征指纹与全部对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与K级特征指纹匹配

级别最高的对应关系中的索引为目标索引，a的初始值为1。

[0020] 采用上述第二种可能的实现方式和第三种可能的实现方式进行数据块的查找，不需要将第一数据块的K级特征指纹与全部的第二数据块的特征指纹进行匹配，能够提高查找效率。

[0021] 结合第二方面、第二方面的第一种可能的实现方式至第三种可能的实现方式中的任一种，在第四种可能的实现方式中，该装置还包括压缩存储单元；压缩存储单元，用于当目标数据块为与第一数据块相同的数据块时，存储指示信息，指示信息用于指示目标数据块的地址；或者，用于当目标数据块为与第一数据块相似的数据块时，基于目标数据块对第一数据块进行相似压缩，并存储压缩后的数据；或者，用于当未确定出目标数据块时，存储第一数据块。

[0022] 该第四种可能的实现方式，在确定目标数据块之后，若目标数据块为与第一数据块相同的数据块，可以通过存储指示信息的方式实现数据去重，数据去重是一种数据无损的冗余数据缩减技术，使得多个相同的数据块在存储系统中只存储一个数据块副本，减少了存储数据所需的资源，节约了成本；若目标数据块为与第一数据块相似的数据块，在确定目标数据块之后，基于目标数据块对第一数据块进行相似压缩，可以减少存储的数据量，提高数据的压缩率，节约存储空间。

[0023] 结合第二方面、第二方面的第一种可能的实现方式至第四种可能的实现方式中的任一种，在第五种可能的实现方式中，该装置还包括执行单元；执行单元，用于计算第一数据块的特征指纹，将第一数据块的特征指纹与第一数据块的索引的对应关系添加至联合查找结构中。

[0024] 第三方面，提供了一种查找数据块的装置，包括：存储器和处理器，存储器用于存储一组代码，处理器用于根据该组代码执行以下动作：获取第一数据块的K级特征指纹，K级特征指纹中的第k级特征指纹的级别高于K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数；根据K级特征指纹和联合查找结构确定目标索引，根据目标索引确定目标数据块，目标数据块为联合查找结构中包括的M个索引对应的M个第二数据块中、与第一数据块相似级别最高的第二数据块，且目标数据块与第一数据块的相似级别大于0，联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，联合查找结构中包含的索引的总个数与第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与第一数据块的相关度。

[0025] 第三方面提供的装置中的各个器件用于执行第一方面提供的方法，因此，该装置的有益效果可以参见上述方法部分的有益效果，在此不再赘述。

[0026] 结合第三方面，在第一种可能的实现方式中，第一数据块和第二数据块的特征指纹包括相似特征指纹，或者，第一数据块和第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

[0027] 结合第三方面的第一种可能的实现方式，在第二种可能的实现方式中，处理器，具体用于：将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配；若匹配成功，确定第一对应关系中的索引为目标索引；若匹配失败、且K级特征指纹大于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合

查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹大的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且K级特征指纹小于第一对应关系中的特征指纹，记录K级特征指纹与第一对应关系中的特征指纹的匹配级别，确定联合查找结构中的与第一对应关系相邻的、且比第一对应关系的特征指纹小的对应关系为新的第一对应关系，将K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与K级特征指纹匹配失败，目标索引为全部的第一对应关系中的与K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为联合查找结构中的第1个对应关系；将K级特征指纹与联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从K级特征指纹中的第1级特征指纹开始，依次与第一对应关系中的相同级别的特征指纹进行匹配，若K级特征指纹中的第K级特征指纹与第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

[0028] 结合第三方面的第一种可能的实现方式，在第三种可能的实现方式中，处理器，具体用于：将K级特征指纹中的第a级特征指纹与联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将K级特征指纹中的第a级特征指纹与全部对应关系中的第a级特征指纹进行匹配，确定与K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

[0029] 采用上述第二种可能的实现方式和第三种可能的实现方式进行数据块的查找，不需要将第一数据块的K级特征指纹与全部的第二数据块的特征指纹进行匹配，能够提高查找效率。

[0030] 结合第三方面、第三方面的第一种可能的实现方式至第三种可能的实现方式中的任一种，在第四种可能的实现方式中，处理器，还用于：当目标数据块为与第一数据块相同的数据块时，存储指示信息，指示信息用于指示目标数据块的地址；当目标数据块为与第一数据块相似的数据块时，基于目标数据块对第一数据块进行相似压缩，并存储压缩后的数据；当未确定出目标数据块时，存储第一数据块。

[0031] 该第四种可能的实现方式，在确定目标数据块之后，若目标数据块为与第一数据块相同的数据块，可以通过存储指示信息的方式实现数据去重，数据去重是一种数据无损的冗余数据缩减技术，使得多个相同的数据块在存储系统中只存储一个数据块副本，减少了存储数据所需的资源，节约了成本；若目标数据块为与第一数据块相似的数据块，在确定目标数据块之后，基于目标数据块对第一数据块进行相似压缩，可以减少存储的数据量，提高数据的压缩率，节约存储空间。

[0032] 结合第三方面、第三方面的第一种可能的实现方式至第四种可能的实现方式中的任一种，在第五种可能的实现方式中，处理器，还用于：计算第一数据块的特征指纹，将第一数据块的特征指纹与第一数据块的索引的对应关系添加至联合查找结构中。

附图说明

[0033] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以

根据这些附图获得其他的附图。

[0034] 图1为现有技术中的一种多级查找结构的组成示意图；

[0035] 图2为本发明实施例提供的一种计算机的硬件架构的组成示意图；

[0036] 图3为本发明实施例提供的一种查找数据块的方法的流程图；

[0037] 图4为本发明实施例提供的一种联合查找结构的组成示意图；

[0038] 图5为本发明实施例提供的一种树型联合查找结构的示意图；

[0039] 图6为本发明实施例提供的一种连续型联合查找结构的示意图；

[0040] 图7为本发明实施例提供的一种查找数据块的装置的组成示意图；

[0041] 图8为本发明实施例提供的又一种查找数据块的装置的组成示意图；

[0042] 图9为本发明实施例提供的再一种查找数据块的装置的组成示意图。

具体实施方式

[0043] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0044] 本发明实施例提供的方法可以应用于数据存储领域，具体可以应用于需要进行多级相似数据查找的压缩存储场景，例如主存场景或备份场景等，另外，还可以应用于图像压缩、版本控制或其它直接压缩效果不好的应用场景。

[0045] 执行本发明实施例提供的方法的设备可以为查找数据的系统或设备，具体可以为计算机等，该计算机的硬件架构组成可以参见图2，包括：输入设备、输出设备、存储器以及中央处理器(Central Processing Unit，简称CPU)等。

[0046] CPU是计算机系统的核心部件，由运算器和控制器组成，运算器主要用于对数据进行加工处理，控制器用于分析指令，并根据指令的要求有序、有目的地向系统的各个部件发出控制信号(具体可参见图2中的细实线箭头的指向)，使整个系统协调一致地工作。存储器能够接收和保存计算机内的数据和程序，还可以根据命令读取已保存的数据和程序，图2中各个器件的数据的流向可以参加图2中粗箭头的指向。根据与CPU的接近程度，存储器可以分为内存储器和外存储器。

[0047] 输入设备用于向计算机输入数据和程序，是用户和计算机进行人机交互的桥梁，主要设备包括：鼠标、键盘、摄像头、扫描仪、光笔以及语音输入装置等，计算机主要通过输入设备获取原始数据和处理这些原始数据的程序。输出设备主要对系统中的数据进行输出，常见的输出设备有：显示器、打印机、绘图仪、影像输出系统以及语音输出系统等。

[0048] 为了使本发明实施例描述的内容更加的清楚，此处对本发明实施例中出现的部分名词进行解释，具体如下：

[0049] 数据块：数据块是一组或几组按顺序连续排列在一起的记录。

[0050] 数据块的特征指纹：采用预设算法(例如，局部敏感哈希(Local Sensitive Hash，简称LSH)算法)对数据块中包含的记录进行计算得到的用于确定与该数据块相关度不为0的数据块的参数。

[0051] 数据块的相同特征指纹：用于查找与该数据块相同(即相关度为1)的数据块的特

征指纹，任意两个不同的数据块的相同特征指纹均不同。

[0052] 数据块的相似特征指纹：用于查找与该数据块相关度大于0小于1的数据块的特征指纹。

[0053] 数据块的特征指纹的级别：对数据块的特征指纹进行的分级，级别越高的特征指纹用于确定与该数据块相关度越高的数据块。

[0054] 联合查找结构：根据已存储的数据块的特征指纹和该数据块的索引组成的用于为待存储的数据块确定已存储的数据块中的与待存储的数据块相关度不为0的数据块的查找结构。

[0055] 相似级别：用于评定两个数据块之间相关度的参数，相似级别越高，这两个数据块之间相关度越高。

[0056] 匹配级别：一个数据块的特征指纹与另一个数据块的特征指纹能够匹配到的级别数，该级别数即两个数据块的相似级别。

[0057] 本发明实施例提供了一种查找数据块的方法，如图3所示，包括：

[0058] 301、获取第一数据块的K级特征指纹，所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数。

[0059] 可选的，在步骤301之前，该方法还包括：获取第一数据块；该情况下，步骤301在具体实现时可以为：计算第一数据块的K级特征指纹。第一数据块可以为由用户上传的数据块，也可以是接收到的其他设备发送的数据块。K的值可以是默认的，也可以是发送第一数据块的设备指示的。

[0060] 具体的，第一数据块可以为待存储的数据块，可以采用LSH算法计算第一数据块的特征指纹。

[0061] 302、根据所述K级特征指纹和联合查找结构确定目标索引，根据所述目标索引确定目标数据块，所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块，且所述目标数据块与所述第一数据块的相似级别大于0，所述联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，所述联合查找结构中包含的索引的总个数与所述第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与所述第一数据块的相关度。

[0062] 其中，第一数据块和第二数据块的相似级别的确定方法可以为：当一个第二数据块的第k′级特征指纹与所述第一数据块的第k′级特征指纹相同、且该第二数据块的第k′+1级特征指纹与所述第一数据块的第k′+1级特征指纹不同时，该第二数据块与所述第一数据块的相似级别为k′，当k′的值越大时，该第二数据块与所述第一数据块的相关度越高，k′为大于等于0小于K的整数。

[0063] 示例性的，当第二数据块与第一数据块的第1级特征指纹相同，第2级特征指纹不同时，第二数据块与第一数据块的相似级别为1，当第二数据块与第一数据块的第1-4级特征指纹均相同，第5级特征指纹不同时，第二数据块与第一数据块的相似级别为4。

[0064] 具体的，第二数据块的索引用于唯一确定该第二数据块，第二数据块的索引具体可以为第二数据块的指针，该指针可以指向该第二数据块的地址。

[0065] 为了便于理解本发明，此处以人为例，对特征指纹的级别进行一个说明，在描述一个人的居住地时，通常可以通过描述这个人居住的国家、居住的省、居住的市、居住的区以及居住的街道的顺序对这个人的居住地进行详细的定位，其中，居住的国家、居住的省、居住的市、居住的区和居住的街道均可以认为是这个人的特征指纹，且这些特征指纹的级别依次增大。由此可以理解，当两个数据块的第k′+1级特征指纹相同，则该两个数据块的第k′级特征指纹必然相同，这就好比居住在同一个城市的俩个人一定是居住在同一个省的。

[0066] 可选的，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹，或者，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹和相同特征指纹。当第一数据块和第二数据块的特征指纹既包括相似特征指纹也包括相同特征指纹时，本发明实施例提供的方法既可以查找与第一数据块相似的第二数据块，也可以查找与第一数据块相同的第二数据块。

[0067] 示例性的，如图4所示，图4示出了一种联合查找结构以及联合查找结构中的索引与第二数据块的对应关系，联合查找结构包括M个对应关系，若M个第二数据块均有N级特征指纹，第m个对应关系为第m个第二数据块Bm的N级特征指纹{Fp1-Bm，Fp2-Bm，…，FpN-Bm}与第m个第二数据块Bm的索引I-Bm的对应关系，联合查找结构中包含的索引的总个数与第二数据块的总个数相同，一个第二数据块的全部的特征指纹对应唯一一个该第二数据块的索引。

[0068] 本发明实施例提供的联合查找结构，若对某个第二数据块进行了更新，直接更新联合查找结构中的包含该第二数据块的索引的对应关系中的数据即可，维护简单、方便。而现有技术中的多级查找结构，一旦有数据更新，则需要更新全部的查找结构，维护不便。需要说明的是，M个第二数据块的特征指纹的级别数一般情况下相同，但是也可以不同，例如，当M＝2时，第一个第二数据块的特征指纹的级别可以有3级，第二个第二数据块的特征指纹的级别可以有5级。需要说明的是，确定第一数据块和第二数据块的同一级的特征指纹以及不同的第二数据块的同一级的特征指纹的方法相同。

[0069] 可选的，步骤302在具体实现时，可以采用以下方式中的任意一种方式实现。

[0070] 方式一：将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配；

[0071] 若匹配成功，确定所述第一对应关系中的索引为目标索引；

[0072] 若匹配失败、且所述K级特征指纹大于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹大的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且所述K级特征指纹小于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹小的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应关系均与所述K级特征指纹匹配失败，所述目标索引为所述全部的第一对应关系中的与所述K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为所述联合查找结构中的第1个对应关系；

[0073] 将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从所述K级特征指纹中的第1级特征指纹开始，依次与所述第一对应关系中的相同级别的特征指纹进行匹配，若所述K级特征指纹中的第K级特征指纹与所述第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

[0074] 上述过程中，将两个特征指纹进行匹配即判断这两个特征指纹是否相同，若相同，则这两个特征指纹匹配成功，否则，匹配失败。

[0075] 其中，当第一对应关系中的第k′级特征指纹与K级特征指纹中的第k′级特征指纹匹配成功、且该第一对应关系中的第k′+1级特征指纹与K级特征指纹中的第k′+1级特征指纹匹配失败时，该第一对应关系与K级特征指纹的匹配级别为k′，由此可知，该第一对应关系与K级特征指纹的匹配级别即该第一对应关系中的索引对应的第二数据块与第一数据块的相似级别。

[0076] 其中，当K级特征指纹中的前x个特征指纹与第一对应关系中的前x个特征指纹相同、且K级特征指纹中的第x+1个特征指纹大于第一对应关系中的第x+1个特征指纹，则K级特征指纹大于第一对应关系中的特征指纹，当K级特征指纹中的前y个特征指纹与第一对应关系中的前y个特征指纹相同、且K级特征指纹中的第y+1个特征指纹小于第一对应关系中的第y+1个特征指纹，则K级特征指纹的特征指纹小于第一对应关系中的特征指纹，x、y均为大于等于0的整数，本发明实施例中，特征指纹的大小为采用预设的算法对同一级的各个特征指纹进行计算得到的结果，用于对同一级的各个指纹进行比较，并没有其他的含义。

[0077] 具体的，在执行方式一时，为了更加方便的对数据块进行查找，可以基于树型结构进行联合查找结构的设计和实现。

[0078] 示例性的，以K＝2且第二数据块有2级特征指纹为例，假设有7个第二数据块，联合查找结构中包括7个对应关系，7个对应关系中的2级特征指纹分别为：[L2,H5]、[L1,H2]、[L4,H4]、[L5,H6]、[L4,H3]、[L2,H7]和[L4,H1]，其中，L表示相似指纹，H表示相同指纹，L和H后的数字代表指纹的大小，则树型联合查找结构如图5所示(未示出特征指纹与索引的对应关系)。

[0079] 若第一数据块的K级特征指纹为[L4,H3]，则采用上述可选的方法查找数据块的过程为：将第一数据块的第1级特征指纹L4与7个对应关系中的第1个对应关系中的第1级特征指纹L2进行匹配，匹配不成功、且L4大于L2，匹配级别为0，确定新的第一对应关系中的特征指纹为[L4,H4]，将第一数据块的第1级特征指纹L4与[L4,H4]中的L4进行匹配，匹配成功，将第一数据块的第2级特征指纹H3与[L4,H4]中的H4进行匹配，匹配不成功、且H3小于H4，匹配级别为1，确定新的第一对应关系中的特征指纹为[L4,H3]，将第一数据块的第1级特征指纹L4与[L4,H3]中的L4进行匹配，匹配成功，将第一数据块的第2级特征指纹H3与[L4,H3]中的H3进行匹配，匹配成功，确定该第一对应关系中的索引为目标索引。具体的上述过程的匹配流程可参加图5中的箭头1和箭头2所示的流程。

[0080] 基于图5所述的示例，若第一数据块的K级特征指纹为[L4,H1]，则采用上述可选的方法查找数据块的过程为：将第一数据块的第1级特征指纹L4与7个对应关系中的第1个对应关系的第1级特征指纹L2进行匹配，匹配不成功、且L4大于L2，匹配级别为0，确定新的第一对应关系中的特征指纹为[L4,H4]，将第一数据块的第1级特征指纹L4与[L4,H4]中的L4进行匹配，匹配成功，将第一数据块的第2级特征指纹H1与[L4,H4]中的H4进行匹配，匹配不

成功、且H1小于H4，匹配级别为1，确定新的第一对应关系中的特征指纹为[L4,H3]，将第一数据块的第1级特征指纹L4与[L4,H3]中的L4进行匹配，匹配成功，将第一数据块的第2级特征指纹H1与[L4,H3]中的H3进行匹配，匹配不成功、且H1小于H3，匹配级别为1，确定新的第一对应关系中的特征指纹为[L2,H7]，将第一数据块的第1级特征指纹L4与[L2,H7]中的L2进行匹配，匹配不成功、且L4大于L2，确定新的第一对应关系中的特征指纹为[L4,H1]，将第一数据块的第1级特征指纹L4与[L4,H1]中的L4进行匹配，匹配成功，将第一数据块的第2级特征指纹H1与[L4,H1]中的H1进行匹配，匹配成功，确定该第一对应关系中的索引为目标索引。具体的上述过程的匹配流程可参加图5中的箭头1、箭头2、箭头3和箭头4所示的流程。

[0081] 方式二：将所述K级特征指纹中的第a级特征指纹与所述联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将所述K级特征指纹中的第a级特征指纹与所述全部对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与所述K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

[0082] 具体的，在执行方式二时，为了更加方便的对数据块进行查找，可以基于数组或链表等连续型结构进行联合查找结构的设计和实现。

[0083] 示例性的，如图6所示，第1-10个数据块的3级特征指纹分别为：[L1,S1,H2]、[L2,S2,H7]、[L3,S4,H4]、[L2,S2,H21]、[L3,S4,H6]、[L1,S3,H5]、[L1,S3,H18]、[L2,S2,H1]、[L1,S1,H3]和[L3,S4,H8]。则这10个数据块的特征指纹在存储时可以存储成连续型联合查找结构，该连续型联合查找结构中包括10个对应关系(未示出对应关系中的索引)，具体如图6所示，其中，第2级特征指纹相同的所有对应关系连续存储，第1级特征指纹相同的所有对应关系也连续存储。

[0084] 方式二在具体实现时，若第一数据块包括3级特征指纹，具体为[L2,S2,H1]，则可以将第一数据块的3级特征指纹中的第1级特征指纹L2与10个对应关系的第1级特征指纹进行匹配，确定与L2相同的全部的对应关系中的特征指纹为[L2,S2,H7]、[L2,S2,H21]、[L2,S2,H1]，将第一数据块的3级特征指纹中的第2级特征指纹S2与[L2,S2,H7]、[L2,S2,H21]、[L2,S2,H1]中的第2级特征指纹进行匹配，确定与S2匹配的全部对应关系中的特征指纹为[L2,S2,H7]、[L2,S2,H21]、[L2,S2,H1]，将第一数据块的3级特征指纹中的第3级特征指纹H1与[L2,S2,H7]、[L2,S2,H21]、[L2,S2,H1]中的第3级特征指纹进行匹配，确定与H1匹配的全部对应关系中的特征指纹为[L2,S2,H1]，则确定第7个对应关系中的索引为目标索引。

[0085] 通过采用方式一和方式二的方法进行数据块的查找，不需要将第一数据块的K级特征指纹与全部的第二数据块的特征指纹进行匹配，能够提高查找效率。

[0086] 可选的，在步骤302之后，所述方法还包括：

[0087] 若所述目标数据块为与所述第一数据块相同的数据块，存储指示信息，所述指示信息用于指示所述目标数据块的地址；

[0088] 若所述目标数据块为与所述第一数据块相似的数据块，基于所述目标数据块对所述第一数据块进行相似压缩，并存储压缩后的数据；

[0089] 若未确定出所述目标数据块，存储所述第一数据块。

[0090] 该可选的方法，在确定目标数据块之后，若目标数据块为与第一数据块相同的数

据块，可以通过存储指示信息的方式实现数据去重，数据去重是一种数据无损的冗余数据缩减技术，使得多个相同的数据块在存储系统中只存储一个数据块副本，减少了存储数据所需的资源，节约了成本；若目标数据块为与第一数据块相似的数据块，在确定目标数据块之后，基于目标数据块对第一数据块进行相似压缩，可以减少存储的数据量，提高数据的压缩率，节约存储空间；若未确定出目标数据块，说明M个第二数据块中没有与第一数据块相同或相似的数据块，则直接存储第一数据块。

[0091] 可选的，在步骤302之后，所述方法还包括：计算所述第一数据块的特征指纹，将所述第一数据块的特征指纹与所述第一数据块的索引的对应关系添加至所述联合查找结构中。

[0092] 本发明实施例提供的方法可以通过联合查找结构对相同数据或相似数据进行查找，由于联合查找结构中包含的索引的总个数与第二数据块的总个数相同，一个第二数据块的全部的特征指纹对应唯一一个该第二数据块的索引，因此，与现有技术中的方案相比，可以大大的减少数据的冗余。

[0093] 本发明实施例还提供了一种查找数据块的装置70，如图7所示，包括：

[0094] 获取单元701，用于获取第一数据块的K级特征指纹，所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数；

[0095] 确定单元702，用于根据所述K级特征指纹和联合查找结构确定目标索引，根据所述目标索引确定目标数据块，所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块，且所述目标数据块与所述第一数据块的相似级别大于0，所述联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，所述联合查找结构中包含的索引的总个数与第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与所述第一数据块的相关度。

[0096] 可选的，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹，或者，所述第一数据块和所述第二数据块的特征指纹包括相似特征指纹和相同特征指纹。

[0097] 可选的，所述确定单元702，具体用于：

[0098] 将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配；

[0099] 若匹配成功，确定所述第一对应关系中的索引为目标索引；

[0100] 若匹配失败、且所述K级特征指纹大于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹大的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系的特征指纹继续进行匹配；若匹配失败、且所述K级特征指纹小于所述第一对应关系中的特征指纹，记录所述K级特征指纹与所述第一对应关系中的特征指纹的匹配级别，确定所述联合查找结构中的与所述第一对应关系相邻的、且比所述第一对应关系的特征指纹小的对应关系为新的第一对应关系，将所述K级特征指纹与新的第一对应关系继续进行匹配，直至确定出目标索引，若全部的第一对应

关系均与所述K级特征指纹匹配失败，所述目标索引为所述全部的第一对应关系中的与所述K级特征指纹匹配级别最高的第一对应关系中的索引，初始的第一对应关系为所述联合查找结构中的第1个对应关系；

[0101] 将所述K级特征指纹与所述联合查找结构中的第一对应关系中的特征指纹进行匹配，包括：从所述K级特征指纹中的第1级特征指纹开始，依次与所述第一对应关系中的相同级别的特征指纹进行匹配，若所述K级特征指纹中的第K级特征指纹与所述第一对应关系中的第K级特征指纹匹配成功，则匹配成功，否则，匹配失败。

[0102] 可选的，所述确定单元702，具体用于：

[0103] 将所述K级特征指纹中的第a级特征指纹与所述联合查找结构中的M个对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，令a＝a+1，将所述K级特征指纹中的第a级特征指纹与所述全部对应关系中的第a级特征指纹进行匹配，确定与所述K级特征指纹中的第a级特征指纹匹配的全部对应关系，直至确定出与所述K级特征指纹匹配级别最高的对应关系中的索引为目标索引，a的初始值为1。

[0104] 可选的，如图8所示，所述装置70还包括压缩存储单元703；

[0105] 所述压缩存储单元703，用于当所述目标数据块为与所述第一数据块相同的数据块时，存储指示信息，所述指示信息用于指示所述目标数据块的地址；或者，用于当所述目标数据块为与所述第一数据块相似的数据块时，基于所述目标数据块对所述第一数据块进行相似压缩，并存储压缩后的数据；或者，用于当未确定出所述目标数据块时，存储所述第一数据块。

[0106] 可选的，如图8所示，所述装置70还包括执行单元704；

[0107] 所述执行单元704，用于计算所述第一数据块的特征指纹，将所述第一数据块的特征指纹与所述第一数据块的索引的对应关系添加至所述联合查找结构中。

[0108] 本发明实施例提供的装置70中的各个单元用于执行上述方法，因此，该装置70的有益效果可以参见上述方法部分的有益效果，在此不再赘述。

[0109] 本发明实施例还提供了一种查找数据块的装置90，如图9所示，包括：存储器901和处理器902，存储器901用于存储一组代码，处理器902用于根据该组代码执行上述图3中所示的方法。

[0110] 其中，查找数据块的装置中的各个功能单元可以以硬件形式内嵌于或独立于查找数据块的装置的处理器中，也可以以软件形式存储于查找数据块的装置的处理器中，以便于处理器调用执行以上各个单元对应的操作。上述处理器可以为CPU、通用处理器、数字信号处理器(Digital Signal Processor，简称DSP)、特定集成电路(Application Specific Integrated Circuit，简称ASIC)、现场可编程门阵列(Field Programmable Gate Array，简称FPGA)或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本发明公开内容所描述的各种示例性的逻辑方框，模块和电路。处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，DSP和微处理器的组合等等。

[0111] 本发明实施例提供的装置90中的各个器件用于执行上述方法，因此，该装置90的有益效果可以参见上述方法部分的有益效果，在此不再赘述。结合本发明公开内容所描述的方法或者算法的步骤可以以硬件的方式来实现，也可以是由处理器执行软件指令的方式

来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于随机存取存储器(Random Access Memory，简称RAM)、闪存、只读存储器(Read Only Memory，简称ROM)、可擦除可编程只读存储器(Erasable Programmable ROM，简称EPROM)、电可擦可编程只读存储器(Electrically EPROM，简称EEPROM)、寄存器、硬盘、移动硬盘、只读光盘(CD-ROM)或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于ASIC中。

[0112] 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。计算机可读介质包括计算机存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0113] 以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

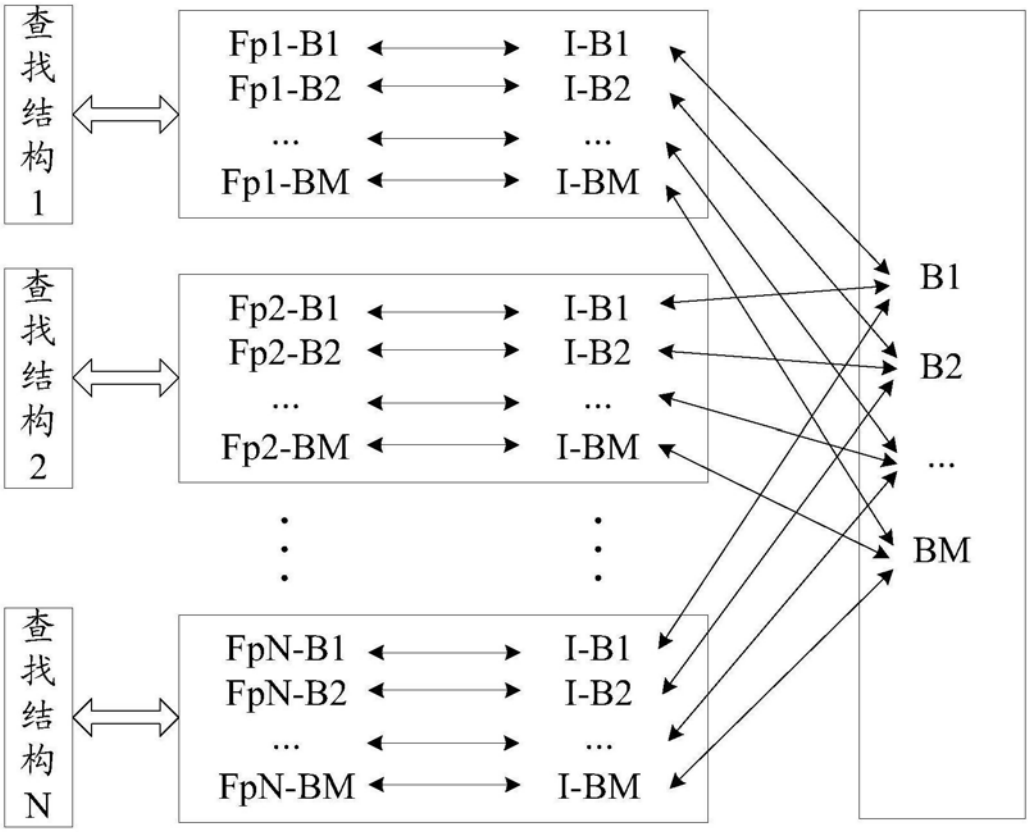

查找结构1
查找结构2
查找结构N
Fp1-B1
Fp1-B2
Fp1-BM
Fp2-B1
Fp2-B2
Fp2-BM
FpN-B1
FpN-B2
FpN-BM
I-B1
I-B2
I-BM
B1
B2
BM

图1

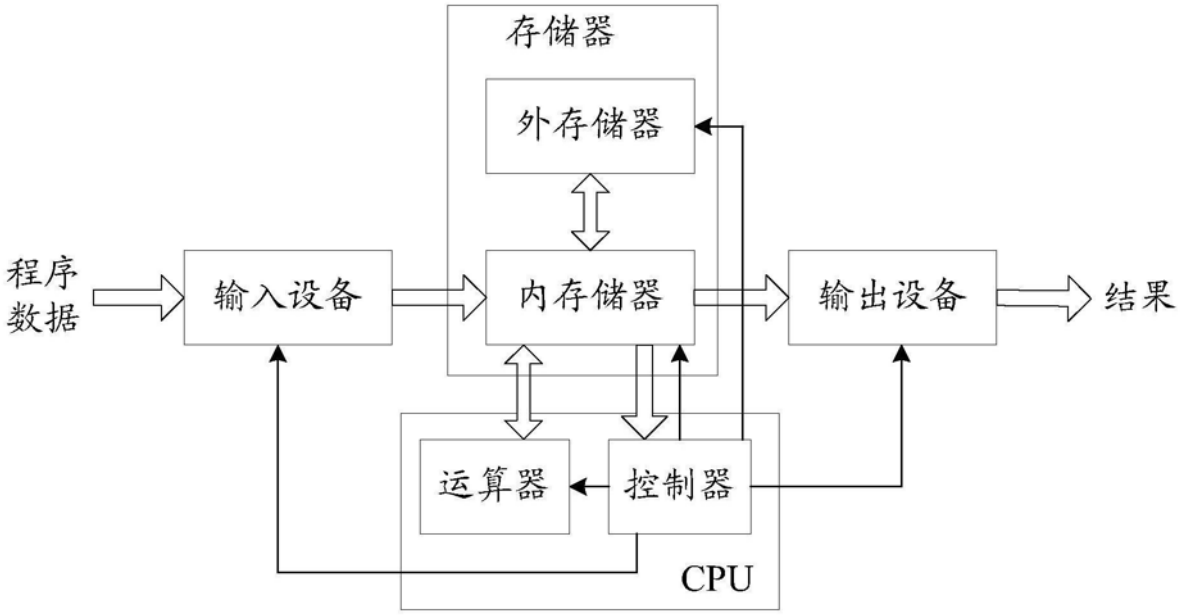

存储器
外存储器
程序数据
输入设备
内存储器
输出设备
结果
运算器
控制器
CPU

图2

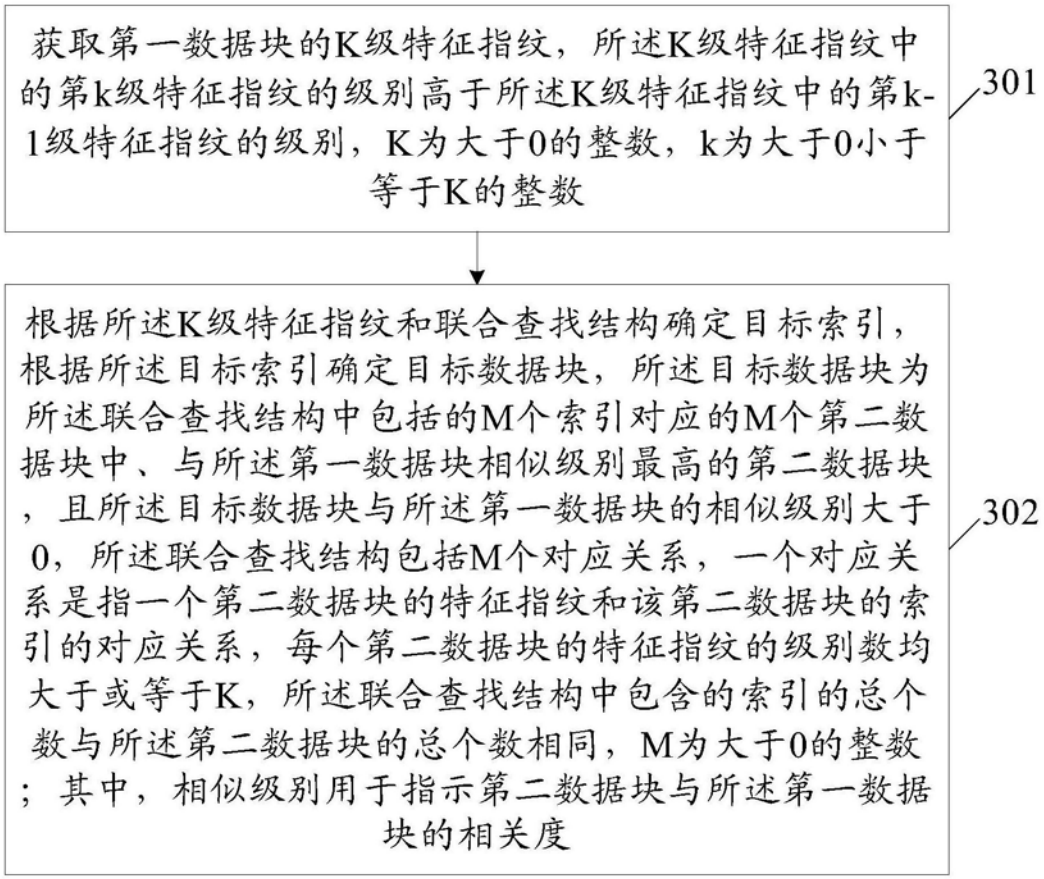

获取第一数据块的K级特征指纹，所述K级特征指纹中的第k级特征指纹的级别高于所述K级特征指纹中的第k-1级特征指纹的级别，K为大于0的整数，k为大于0小于等于K的整数
301
根据所述K级特征指纹和联合查找结构确定目标索引，根据所述目标索引确定目标数据块，所述目标数据块为所述联合查找结构中包括的M个索引对应的M个第二数据块中、与所述第一数据块相似级别最高的第二数据块，且所述目标数据块与所述第一数据块的相似级别大于0，所述联合查找结构包括M个对应关系，一个对应关系是指一个第二数据块的特征指纹和该第二数据块的索引的对应关系，每个第二数据块的特征指纹的级别数均大于或等于K，所述联合查找结构中包含的索引的总个数与所述第二数据块的总个数相同，M为大于0的整数；其中，相似级别用于指示第二数据块与所述第一数据块的相关度
302

图3

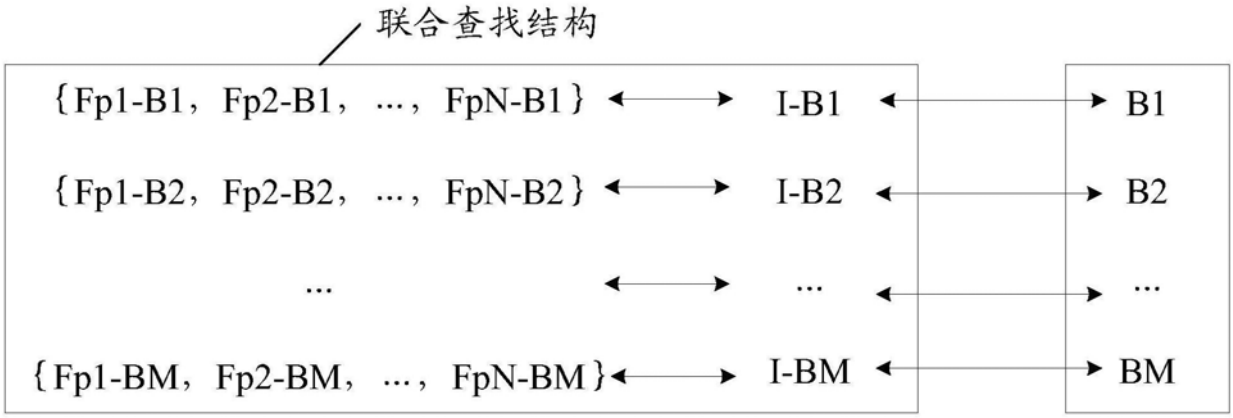

联合查找结构
{Fp1-B1，Fp2-B1，…，FpN-B1}
I-B1
B1
{Fp1-B2，Fp2-B2，…，FpN-B2}
I-B2
B2
…
…
…
{Fp1-BM，Fp2-BM，…，FpN-BM}
I-BM
BM

图4

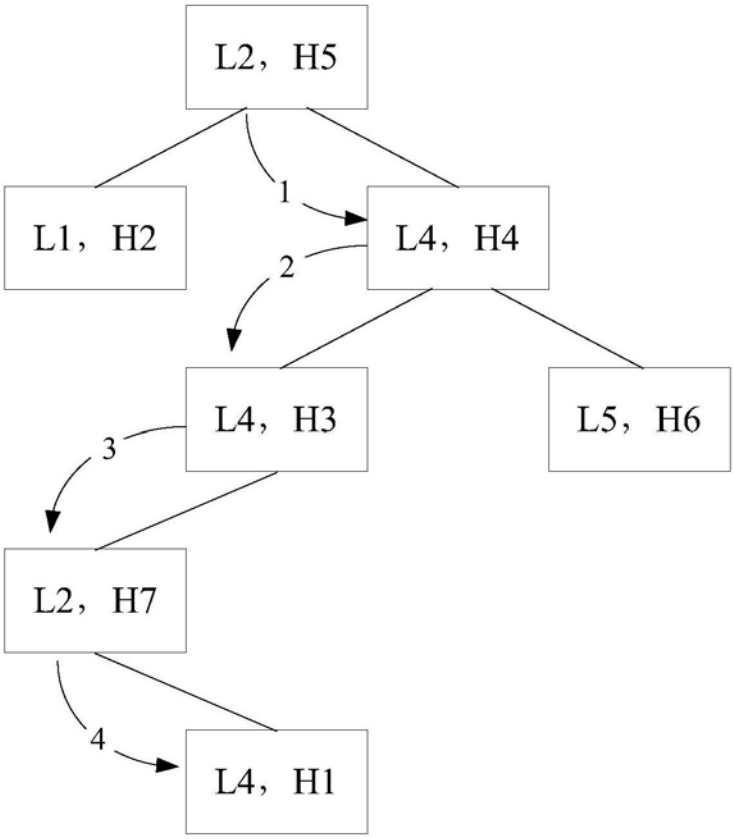


图5

L1	S1	H2
		H3
	S3	H18
		H5
L2	S2	H7
		H21
		H1
L3	S4	H6
		H4
		H8

图6

图7


获取单元701
确定单元702
压缩存储单元703
执行单元704
查找数据块的装置70

图8


存储器901
处理器902
查找数据块的装置90

图9