

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 7 月 21 日 (21.07.2022)



(10) 国际公布号
WO 2022/152086 A1

(51) 国际专利分类号:
G06F 16/2455 (2019.01)

(21) 国际申请号: PCT/CN2022/071079

(22) 国际申请日: 2022 年 1 月 10 日 (10.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110057914.1 2021 年 1 月 15 日 (15.01.2021) CN

(71) 申请人: 华为云计算技术有限公司 (HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.)
[CN/CN]; 中国贵州省贵安新区黔中大道交兴功
路华为云数据中心, Guizhou 550025 (CN)。

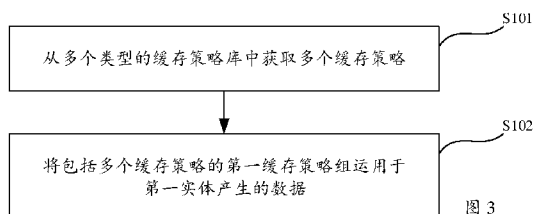
(72) 发明人: 郭 畅 (GUO, Chang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** DATA CACHING METHOD AND APPARATUS, AND DEVICE AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 数据缓存方法、装置、设备及计算机可读存储介质



S101 Acquire a plurality of caching policies from a plurality of types of caching policy libraries
S102 Apply a first caching policy group, which comprises the plurality of caching policies, to data generated by a first entity

图 3

(57) **Abstract:** A data caching method and apparatus, and a device and a computer-readable storage medium. The method comprises: acquiring a plurality of caching policies from a plurality of types of caching policy libraries (S101), wherein at least one caching policy comprised in each type of caching policy libraries is of the same type, the types of the caching policy libraries comprise at least one of a filtering type, a pre-fetching type, a replacement type and a victim cache type, a caching policy of the filtering type is used for filtering data, a caching policy of the pre-fetching type is used for pre-fetching data, a caching policy of the replacement type is used for eliminating data from a cache, and a caching policy of the victim cache type is used for processing the data eliminated from the cache; and applying a first caching policy group, which comprises the plurality of caching policies, to data generated by a first entity (S102). By using the above-mentioned method, a caching policy applied to data generated by a first entity can be flexibly selected, so as to improve the caching effect of a first caching policy group on the data generated by the first entity.

(57) **摘要:** 一种数据缓存方法、装置、设备及计算机可读存储介质, 其中, 该方法包括: 从多个类型的缓存策略库中获取多个缓存策略 (S101), 每个类型的缓存策略库包括的至少一个缓存策略的类型相同, 其中, 缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个, 过滤类型的缓存策略用于过滤数据, 预取类型的缓存策略用于预取数据, 替换类型的缓存策略用于从缓存中淘汰数据, 牺牲缓存类型的缓存策略用于处理从缓存中淘汰的数据; 将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据 (S102)。利用上述方法可以灵活地选取运用于第一实体产生的数据的缓存策略, 从而提高第一缓存策略组对第一实体产生的数据的缓存效果。

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

数据缓存方法、装置、设备及计算机可读存储介质

本申请要求于 2021 年 1 月 15 日提交中国国家知识产权局、申请号为 202110057914.1、发明名称为“数据缓存方法、装置、设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及缓存技术领域，尤其涉及一种数据缓存方法、装置、设备及计算机可读存储介质。

背景技术

随着缓存技术的快速发展，用户获取数据的速度也越来越快。其中，影响用户获取数据的速度的一個关键因素是缓存策略，一个合适的缓存策略能够保证缓存的有效性，从而提高用户获取数据的速度。但是，在实际应用过程中，由于业务逻辑的不同，不同用户的数据访问模式也可能不同，那么适用于处理这些用户产生的数据的缓存策略也会有所差异。因此，为了适应不同用户的数据访问模式，构建出灵活且适应能力强的缓存策略是当前缓存技术领域中的一个亟需解决的问题。

发明内容

本申请公开了一种数据缓存方法、装置、设备及计算机可读存储介质，能够构建出灵活度高且适应能力强的缓存策略组，并且利用该缓存策略组进行数据缓存时，可以提高缓存的有效性，从而提高数据读取的速度。

第一方面，本申请提供了一种数据缓存方法，该方法包括如下步骤：

从多个类型的缓存策略库中获取多个缓存策略，每个类型的缓存策略库包括的至少一个缓存策略的类型相同，其中，缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个，过滤类型的缓存策略用于过滤数据，预取类型的缓存策略用于预取数据，替换类型的缓存策略用于从缓存中淘汰数据，牺牲缓存类型的缓存策略用于处理从缓存中淘汰的数据；

将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

实施第一方面所描述的方法，用户或者设备可以自行选择运用于第一实体产生的数据的缓存策略，在这一过程中，用户或设备可以根据实际需求灵活地选择缓存策略，不仅可以使得第一缓存策略组具有更多的灵活性，还可以提高第一缓存策略组对第一实体产生的数据的缓存效果。

在一种可能的实现方式中，将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据，包括：根据第一实体产生的数据的访问记录对第一缓存策略组进行优化，得到第二缓存策略组；将第二缓存策略组运用于第一实体产生的数据。

上述实现方式中，由于第二缓存策略组是根据第一实体产生的数据的访问记录对第一缓存策略组进行优化得到的，因此相较于第一缓存策略组，第二缓存策略组更适合处理第一实

体产生的数据，也就是说，利用第二缓存策略组对第一实体产生的数据进行缓存可以达到更好的缓存效果，从而提高第一实体获取数据的速度。

在一种可能的实现方式中，缓存策略库的类型还包括专有类型，专有类型的缓存策略为用户设置的缓存策略。

上述实现方式中，用户可以自行设置所需的缓存策略，那么，第一缓存策略组可以具有更高的灵活性。

在一种可能的实现方式中，上述方法还包括：根据第二实体产生的数据的访问记录对第一缓存策略组进行优化，得到第三缓存策略组；将第三缓存策略组运用于第二实体产生的数据。

上述实现方式中，第一缓存策略组还可以用于对第二实体产生的数据进行处理，其中，第一实体产生的数据与第二实体产生的数据不同，因此，第一缓存策略组能够适用于不同实体产生的数据，具有良好的适应能力。

在一种可能的实现方式中，第一缓存策略组中的多个缓存策略是按照预设顺序排布的，当多个缓存策略包括专有类型的缓存策略时，专有类型的缓存策略的位置由用户设置。

上述实现方式中，用户还可以自行设置缓存策略的位置，那么，用户可以根据自身需求来设置第一缓存策略组，从而使得第一缓存策略组可以具有更高的灵活性。

在一种可能的实现方式中，在将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据之前，上述方法还包括：确定第一缓存策略组的合法性。

可以理解的，由于第一缓存策略组包括多个缓存策略，而多个缓存策略之间可能存在冲突，因此为了避免这种情况，在运用第一缓存策略组进行数据缓存之前，需要先对第一缓存策略组进行合法性检查。当第一缓存策略组合法的情况下，可以将其运用于第一实体产生的数据，当第一缓存策略组不合法的情况下，需要将其进一步调整为合法的缓存策略组，再运用于第一实体产生的数据。

在一种可能的实现方式中，多个缓存策略中的每个缓存策略对应一个策略属性集合，确定第一缓存策略组的合法性，包括：根据多个缓存策略对应的多个策略属性集合，确定第一缓存策略组的合法性；其中，一个缓存策略对应的策略属性集合包括第一属性、第二属性中的至少一个，第一属性用于确定第一缓存策略组中是否存在与上述一个缓存策略冲突的缓存策略，第二属性用于确定第一缓存策略组中能否包括多个上述一个缓存策略。

上述实现方式中，通过为每个缓存策略设置策略属性集合，可以更加方便地对第一缓存策略组的合法性进行确定。

在一种可能的实现方式中，根据第一实体产生的数据的访问记录对第一缓存策略组进行优化，得到第二缓存策略组，包括：在第一缓存策略组合法的情况下，根据第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对第一缓存策略组中的每个缓存策略进行迭代优化，从而得到第二缓存策略组。

上述实现方式中，利用启发式算法或机器学习算法来优化第一缓存策略组中的每个缓存策略，可以提高优化速度，从而可以更快的获得第二缓存策略组。

第二方面，本申请提供了一种数据缓存装置，所述装置包括：

获取单元，用于从多个类型的缓存策略库中获取多个缓存策略，每个类型的缓存策略库包括的至少一个缓存策略的类型相同，其中，缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个，过滤类型的缓存策略用于过滤数据，预取类型的缓存策略用于预取数据，替换类型的缓存策略用于从缓存中淘汰数据，牺牲缓存类型的缓存策

略用于处理从缓存中淘汰的数据；

缓存单元，用于将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

在一种可能的实现方式中，缓存单元具体用于：根据第一实体产生的数据的访问记录对第一缓存策略组进行优化，得到第二缓存策略组；将第二缓存策略组运用于第一实体产生的数据。

在一种可能的实现方式中，缓存策略库的类型还包括专有类型，专有类型的缓存策略为用户设置的缓存策略。

在一种可能的实现方式中，缓存单元还用于：根据第二实体产生的数据的访问记录对第一缓存策略组进行优化，得到第三缓存策略组；将第三缓存策略组运用于第二实体产生的数据。

在一种可能的实现方式中，第一缓存策略组中的多个缓存策略是按照预设顺序排布的，当多个缓存策略包括专有类型的缓存策略时，专有类型的缓存策略的位置由用户设置。

在一种可能的实现方式中，上述装置还包括确定单元，确定单元用于：确定第一缓存策略组的合法性。

在一种可能的实现方式中，多个缓存策略中的每个缓存策略对应一个策略属性集合，确定单元具体用于：根据多个缓存策略对应的多个策略属性集合，确定第一缓存策略组的合法性；其中，一个缓存策略对应的策略属性集合包括第一属性、第二属性中的至少一个，第一属性用于确定第一缓存策略组中是否存在与上述一个缓存策略冲突的缓存策略，第二属性用于确定第一缓存策略组中能否包括多个上述一个缓存策略。

在一种可能的实现方式中，缓存单元具体用于：在第一缓存策略组合法的情况下，根据第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对第一缓存策略组中的每个缓存策略进行迭代优化，从而得到第二缓存策略组。

第三方面，本申请提供了一种缓存设备，缓存设备包括处理器和存储器，处理器执行存储器中的代码以实现第一方面所描述的部分或全部步骤。

第四方面，本申请提供了一种计算机可读存储介质，存储有计算机指令，计算机指令用于实现第一方面所描述的部分或全部步骤。

第五方面，本申请提供了一种计算机程序产品，包括计算机程序，当计算机程序被计算设备读取并执行时，实现如第一方面所描述的部分或全部步骤。

附图说明

为了更清楚地说明本申请涉及的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，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。

图 1A 是本申请提供的一种 ARC 策略的示意图；

图 1B 是本申请提供的一种 ARC 策略自适应调整的原理示意图；

图 2A 是本申请提供的一种缓存策略选择界面的示意图；

图 2B 是本申请提供的另一种缓存策略选择界面的示意图；

图 3 是本申请提供的一种数据缓存方法的流程示意图；

图 4A 是本申请提供的一种缓存管理页面的示意图；

图 4B 是本申请提供的另一种缓存管理页面的示意图；

图 5A 是本申请提供的一种第一缓存策略组的示意图；
图 5B 是本申请提供的另一种第一缓存策略组的示意图；
图 6 是本申请提供的一种具体实施例的流程示意图；
图 7 是本申请提供的一种数据缓存装置的结构示意图；
图 8 是本申请提供的一种缓存设备的结构示意图。

具体实施方式

下面将结合附图，对本申请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申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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。

为了便于理解本申请提供的技术方案，首先介绍本申请涉及到的相关概念。

数据访问实体（以下简称为“实体”）：是指访问数据，并对数据存储需求（即缓存需求）有所差异的数据使用者，例如，用户（组）、应用（组）、进程（组）、线程（组）等。本申请中的缓存需求可以理解为实体对存储在缓存中的数据的要求，即缓存中存储哪些数据时，缓存具备有效性。

缓存有效性：是指缓存是否有效，当缓存具备有效性时，实体可以从直接缓存中获取较多的被访问的数据，从而可以提高实体获取数据的速度。一般地，衡量缓存有效性的指标包括：缓存命中率、数据迁移量读、放大倍数。

缓存命中率：当实体访问数据时，如果缓存中存储了该数据，则实体可以从缓存中获取该数据，这就是命中，相反地，如果缓存中没有存储该数据，则实体需要从内存中获取该数据，这就是没有命中。那么， $\text{缓存命中率} = \frac{\text{命中的数量}}{\text{命中的数量} + \text{没有命中的数量}}$ ，可以看出，缓存命中率越高，表示缓存的使用率越高，即被访问的数据大部分都是从缓存中获取到的。

数据迁移量：是指缓存数据的迁移量。为了加快存储系统的性能，通常会根据数据的重要性、访问频率、保留时间、容量、性能等指标，将数据采取不同的存储方式分别存储在不同性能的存储设备上，也就是通过分级存储的方式将不经常被访问的数据自动迁移到存储层次中较低的层次，从而释放出较高成本的存储空间给被频繁访问的数据。那么，当数据迁移量越大时，意味着缓存中存储了大量的不被经常访问的数据，这将增加数据迁移的成本以及存储系统的负担。值得注意的一点时，“数据迁移量”可以具有不同的名称，例如，不同标准或者同一标准的不同版本、不同厂商、不同应用场景对“数据迁移量”可以具有不同的称呼，例如，缓存开销等。

读放大倍数：对于一个数据来说，如果缓存策略判断该数据符合缓存需求，则会从内存中读取该数据，并复制到缓存，这时该数据相当于被读取 2 次。接下来，当缓存已满时，如果缓存策略判断该数据不符合缓存，则会从缓存中淘汰该数据，当该数据再次被访问时，缓存策略又会认为该数据符合缓存需求，那么又会将该数据从内存复制到缓存，这时该数据又被读取了 2 次。那么，当一个数据的读放大倍数越大，意味着该数据被从缓存中反复淘汰，该缓存的缓存效果较差。值得注意的一点是，“读放大倍数”可以具有不同的名称，例如，不同标准或者同一标准的不同版本、不同厂商、不同应用场景对“读放大倍数”可以具有不同的称呼，例如，缓存载入量等。

数据访问模式：是指实体访问数据的方式，例如，最近访问(recency)模式和频繁访问

(frequency)模式,最近访问模式是指实体总是访问最近被访问过的数据,频繁访问模式是指实体总是访问历史访问频率较高的数据。可以理解的,为了提高缓存有效性,缓存中的数据应符合数据访问模式的要求,例如,如果实体采用最近访问模式,那么缓存中应存储最近被访问过的数据,如果实体采用频繁访问模式,那么缓存中应存储历史访问频率较高的数据。换句话说,用于处理实体产生的数据的缓存策略应适应于数据访问模式,这样才能满足缓存需求,从而提高缓存的有效性。

为了便于理解本申请提供的技术方案,首先介绍本申请适用的应用场景:缓存场景。

缓存(cache)是指位于速度相差较大的两种硬件(例如,处理器与内存、内存与硬盘、硬盘与网络等)之间,用于协调二者数据传输速度差异的结构,以达到提高数据读取性能的效果。不难理解,当用户频繁访问一个数据时,如果该数据每次都是从内存中提取到的,那么用户每次访问时都需要等待较长的时间才能获得该数据。而缓存的出现有效地解决了该问题,具体地,将被频繁访问的数据复制到缓存,使得用户后续访问该数据时,都可以直接从缓存中读取该数据,从而提高数据读取的速度。

简单地说,缓存是一个用于存储被频繁访问的数据的存储区。但是,随着时间的推移,被频繁访问的数据可能随之变化,也就是说,前一段时间被频繁访问的数据此时可能不被访问,而前一段间不被频繁访问的数据此时可能被频繁访问,那么如何确定哪些数据属于被频繁访问的数据?另外,当缓存空间被占满时,如果存在新的数据需要被复制到缓存,那么此时缓存中应保留哪些数据,删除哪些数据?因此,需要设计一个缓存策略来管理缓存,以保证缓存的有效性。

值得注意的一点是,“缓存策略”可以具有不同的名称。例如,不同标准或者同一标准的不同版本、不同厂商、不同应用场景对“缓存策略”可以具有不同的称呼,例如,术语“缓存策略”有时也可以被称为“缓存方法”、“缓存算法”等。

对于一个实体来说,合适的缓存策略是提高该实体获取数据的速度的关键,但实际应用中,由于不同实体对应的数据访问模式可能不同,因此不同实体对应的缓存需求也可能不同,那么,适合处理不同实体产生的数据的缓存策略也要有所不同。不难理解,如果对每个实体都设计一个合适的缓存策略,这将耗费大量的资源。因此,为了适应不同用户的数据访问模式,如何构建出灵活且适应能力强的缓存策略,以用于提高缓存的有效性仍然是当前缓存技术领域中的一个亟需解决的问题。

目前,常见的缓存策略方案包括单一自适应策略方案和混合缓存策略方案,具体如下:

(1) 单一自适应策略方案

单一自适应策略方案:是指为不同的实体配置同一个缓存策略,然后在实际运用过程中根据各个实体的数据访问情况对该缓存策略进行改良,使其能够适应于不同实体采用的不同数据访问模式。比如说,Megiddo 和 Modha 提出的具有一定自适应能力的缓存算法(adaptive replacement cache, ARC)就是一种典型的单一自适应策略,这种缓存策略融合了最近最少使用(least recently used, LRU)算法的思想(即,如果一个数据在最近一段时间内未被访问过,那么可以认为这个数据在将来被访问的可能性也很小)和最不经常使用(least frequently used, LFU)算法的思想(即,如果一个数据在最近一段时间内很少被访问,那么可以认为这个数据在将来被访问的可能性也很小)。因此,ARC 策略适合处理采用最近访问模式的实体产生的数据,或者采用频繁访问模式的实体产生的数据。

如图 1A 所示,图 1A 示出了一种 ARC 策略的示意图。从图中可以看出,ARC 策略具体

包括 LRU 链表、LFU 链表、用于存储从 LRU 链表中淘汰的数据的信息的链表 (Ghost LRU 链表)、用于存储从 LFU 链表中淘汰的数据的信息的链表 (Ghost LFU 链表)。其中, LRU 链表和 LFU 链表用于存储数据, 具体地, LRU 链表存储的是最近最多使用的数据, LFU 链表存储的是最近最频繁使用的数据; 而 Ghost LRU 链表和 Ghost LFU 链表中不存储数据, 存储的是数据的信息 (例如, 偏移量(offset))。

在实际应用过程中, ARC 策略会根据这四个链表的命中情况, 动态调整 LRU 链表和 LFU 链表的长度, 从而使得 ARC 策略具有适应能力。具体地, 如果 LRU 链表和 Ghost-LRU 链表的命中率较高, 则增加 LRU 链表的长度, 如果 LFU 链表和 Ghost-LFU 链表的命中率较高, 则增加 LFU 链表的长度。举例说明, 如图 1B 所示, 在 LRU 链表已满的情况下, 如果需要向 LRU 链表中再写入一个数据 A, 则 LRU 链表中最近最少被访问的数据 B 将会被淘汰, 并且数据 B 会被放入 Ghost-LRU 队列。假设一段时间后, 数据 B 被再一次访问, 这时 Ghost-LRU 队列被命中, 在这种情况下, 会将 LRU 链表的长度加 1, 对应的, LFU 链表的长度减 1。

可以看出, ARC 策略的适应能力是基于链表的命中情况, 在 LRU 算法和 LFU 算法之间进行切换而实现的, 因此 ARC 策略仅能适应于最近访问模式和频繁访问模式, 这也就导致 ARC 策略具备的适应能力有限, 难以满足除上述两种数据访问模式对应的缓存需求外的其他缓存需求。类似的, 其他的单一自适应策略也存在灵活度和适应能力有限的问题。总的来说, 单一自适应策略通常存在灵活性和适应能力较差的问题, 从而难以满足众多不同的缓存需求。

(2) 混合缓存策略方案

混合缓存策略方案: 提供有候选策略集合, 对于不同的实体, 可以从候选策略集合中选择相应的缓存策略来处理对应实体产生的数据。目前, 混合缓存策略为云存储、内容分发网络(content delivery network, CDN)等领域中最为常用的缓存策略。

在一种可能的实现方式中, 该方案提供有缓存策略选择界面, 该界面向用户展示了候选策略集合, 使得用户可以自行选择合适的缓存策略。如图 2A 示出的缓存策略选择界面中, 提供有缓存策略的选择选项, 用户可以根据实体对应的数据访问模式选择相应的一个或多个缓存策略。除缓存策略的选择选项外, 缓存策略选择界面还可以提供缓存策略的配置选项, 使得用户可以方便地为指定的文件选择缓存策略, 以图 2B 示出的缓存策略选择界面为例, 用户可以在配置选项中输入“.txt”, 并为文件后缀为“.txt”的文件选择缓存策略 1, 从而实现利用缓存策略 1 来处理文件后缀为“.txt”的文件中的数据。

在另一种可能的实现方式中, 通过分析实体对应的数据访问模式, 然后根据实体的数据访问模式, 从候选策略集合中选择出适应于该实体对应的数据访问模式的缓存策略。

可以看出, 该方案中候选策略集合中的缓存策略越丰富, 越能从中挑选出更合适的缓存策略。但是, 如何构建出包括海量缓存策略的候选策略集合, 本身就是一个难以实现的工作, 这将导致该方案也无法满足众多不同的缓存需求。

为了解决上述问题, 本申请提供了一种数据缓存方法, 能够构建出灵活度高且适应能力强的缓存策略组, 从而提高缓存的有效性。下面, 将结合图 3-图 6 对本申请提供的数据缓存方法进行详细介绍。

首先, 请参见图 3, 图 3 示出了本申请提供的一种数据缓存方法的流程示意图, 该方法包括但不限于如下步骤:

S101: 缓存设备从多个类型的缓存策略库中获取多个缓存策略。

在一具体的实施例, 每个类型的缓存策略库包括的至少一个缓存策略的类型相同, 其

中，缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个。可选的，缓存策略库的类型还包括专有类型。下面对各个类型的缓存策略库及其包括的缓存策略进行介绍：

过滤类型的缓存策略库（以下简称为过滤策略库）包括一个或多个过滤类型的缓存策略（以下简称为过滤策略），过滤策略用于过滤数据，例如，处理器在执行人脸识别的程序时，会产生人脸图像、人脸特征、人脸识别结果等大量的数据，过滤策略可以将人脸图像、人脸特征等数据滤除。本申请中，过滤类型的缓存策略库中包括的过滤策略可以是双边滤波算法(double filter)、布隆滤波算法(bloom filter)等经典的滤波算法，也可以是用户自定义的过滤规则，例如，用户通过自定义的方式将文件名后缀为“.jpg”的文件过滤，本申请对此不作具体限定。

预取类型的缓存策略库（以下简称为预取策略库）包括一个或多个预取类型的缓存策略（以下简称为预取策略），预取策略用于预取数据，具体地说，预取策略用于预测实体将要访问的数据，并将预测得到的数据提前存储到缓存中，那么，当实体访问到这些数据时，便能直接从缓存中获得，从而提高实体访问效率。例如，处理器执行一个循环指令时，预取策略可以根据处理器已经执行该循环指令的次数，预测出下一次执行循环指令时所需的数据，并将处理器下一次执行循环指令时所需的数据提前存储到缓存中。本申请中预取类型的缓存策略库中包括的预取策略可以是预读算法(readahead)，自适应预读算法(adaptive readahead)、智能预取算法(smart prefetcher)等，也可以是用户自定义的预取规则，此处不作具体限定。

替换类型的缓存策略库（以下简称为替换策略库）包括一个或多个替换类型的缓存策略（以下简称为替换策略），替换策略用于从缓存中淘汰数据，本申请中，替换类型的缓存策略库中包括的替换策略可以是LRU、LFU、ARC等，也可以是用户自定义的替换规则，此处不作具体限定。

牺牲缓存类型的缓存策略库（以下简称为牺牲缓存策略库）包括一个或多个牺牲缓存类型的缓存策略（以下简称为牺牲缓存策略），牺牲缓存策略用于处理从缓存中淘汰的数据。可以理解的，由于从缓存中淘汰的数据仍然可能被再次访问，在这种情况下，实体需要重新从内存中获得被淘汰的数据，为了减少这一过程造成的损失，通过牺牲缓存策略可以将被淘汰的数据暂时存储到牺牲缓存(victim cache)中，然后根据后续被访问的概率确定是否从牺牲缓存中淘汰该数据，也就是说，牺牲缓存策略可以将从缓存中被淘汰但后续被访问概率大的数据存储到牺牲缓存中，使得实体再次访问被淘汰的数据时，可以直接从牺牲缓存中获取该数据。

专有类型的缓存策略库（以下简称为专有策略库）包括一个或多个专有类型的缓存策略（以下简称为专有策略），专有策略为用户设置的缓存策略，具体可以是用于根据实体的数据访问行为而专门定制的策略，例如，实体访问数据库时，每次都是按照数据库的行号来提取数据，因此，用户可以设置一个根据数据库的行号来提取相关数据的专有策略。

可选的，多个类型的缓存策略库可以都配置在缓存设备中，也可以都配置在其他电子设备或系统中，还可以部分配置在缓存设备中，部分配置在其他电子设备或系统中，此处不作具体限定。

可选的，从多个类型的缓存策略库获取的多个缓存策略的类型可以都不相同，也可以全都相同，还可以一部分相同，一部分不同，此处不作具体限定。可选的，上述多个缓存策略可以包括多个相同的缓存策略，上述多个缓存策略也可以是多个不同的缓存策略，此处不作具体限定。具体实现中，缓存设备可以通过以下几种方式从多个类型的缓存策略库中获取多

个缓存策略。

方式一，缓存设备获取用户从多个类型的缓存策略库中选取的多个缓存策略。

具体地，多个类型的缓存策略库可以以缓存管理页面的形式向用户进行展示，以图 4A 为例，缓存管理页面上显示了 5 个类型的缓存策略库，包括：过滤策略库、预取策略库、替换策略库、牺牲缓存策略库以及专有策略库，其中，过滤策略库包括 3 个过滤策略，预取策略库包括 4 个预取策略，替换策略库包括 5 个替换策略，牺牲缓存策略库包括 2 个牺牲缓存策略，专有策略库包括 2 个专有策略。那么，用户可以通过缓存管理页面清楚地获知有哪些类型的缓存策略库，以及有哪些缓存策略可供选择，而且，用户还可以在缓存管理页面上进行操作，从而选择出上述多个缓存策略。

可选的，缓存管理页面上还可以显示配置选项，以方便用户可以定义利用缓存策略对哪些数据进行处理、处理后的数据缓存到什么地方以及缓存时间、缓存优先级等。以图 4B 为例，用户通过在缓存管理页面上的配置选项中输入目录 A，则缓存设备会利用选择出多个缓存策略对目录 A 中的数据进行处理。

可选的，用户可以随机地从多个类型的缓存策略库中选取多个缓存策略，也可以通过分析第一实体的数据访问模式（具体可以是第一实体产生的数据的访问记录），从而在多个类型的缓存策略库中选取多个缓存策略，还可以通过其他方式从多个类型的缓存策略库中选取多个缓存策略，本申请对此不作具体限定。

方式二，缓存设备从多个类型的缓存策略库中选取多个缓存策略。

可选的，缓存设备可以随机地从多个类型的缓存策略库中选取多个缓存策略，也可以通过分析第一实体的数据访问模式（具体可以是第一实体产生的数据的访问记录），从而在多个类型的缓存策略库中选取多个缓存策略，还可以根据下发的配置文件从多个类型的缓存策略库中选取多个缓存策略等，本申请对此不作具体限定。其中，配置文件包括以下一个或多个：选取的缓存策略的总数量、选取哪些类型的缓存策略、选取每种类型的缓存策略的数量、具体选取哪个缓存策略。

方式三，缓存设备还可以通过将方式一和方式二进行结合，从而获取多个缓存策略，即一部分缓存策略为用户选取的，另一部分缓存策略为缓存设备选取的。

通过上述方法，用户或者缓存设备可以根据实际需求灵活地选择缓存策略，这使得第一缓存策略组能够更加符合用户的要求，例如，当存储系统中没有配置牺牲缓存时，可以不选择牺牲缓存策略，又例如，当需要将缓存中淘的数据先进行过滤再存入牺牲缓存时，用户可以设置一个专有策略来达到该目的。而且，第一缓存策略组中缓存策略的类型、每种类型的缓存策略的数量等都是可以根据实际情况进行相应调整的，这使得第一缓存策略组具有多种可能性。不难理解，相较于前述内容中提及的单一自适应策略和候选策略集合，利用上述方法可以容易地扩展出更多的缓存策略组，从而提供更多的选择，也就是能够满足更多的缓存需求。

S102：缓存设备将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

在一具体的实施例中，第一缓存策略组中的多个缓存策略是按照预设顺序排布的。当多个缓存策略包括专有策略时，专有策略的位置由用户设置。

更具体地，当多个缓存策略包括过滤策略、预取策略、替换策略、牺牲缓存策略以及专有策略时，预设顺序为：过滤策略排布在预取策略之前，预取策略排布在替换策略之前，替换策略排布在牺牲缓存策略之前，专有策略的位置由用户设置，也就是说，专有策略可以排布于任一策略之前或之后。可以理解的，用户通过指定专用策略的位置，可以使得第一缓存

策略组更加符合缓存需求，从而可以获得更好的缓存效果。

在一具体的实施例中，当上述多个缓存策略中包括两个或两个以上类型相同的缓存策略时，预设顺序还可以定义这些类型相同的缓存策略的顺序。以图 4A 示出的缓存策略为例，预设顺序定义了过滤策略 1 排布在过滤策略 2 之前，过滤策略 3 排布在过滤策略 1 之前。那么，当多个缓存策略包括过滤策略 1、过滤策略 2、预取策略 4、替换策略 3、牺牲缓存策略 1 以及专有策略 2 时，缓存设备可以根据预设顺序将这些缓存策略进行排布，从而得到如图 5A 所示的第一缓存策略组。可选的，预设顺序中也可以不定义类型相同的缓存策略之间的排布顺序，继续以上述例子为例，当预设顺序中没有定义过滤策略 1 和过滤策略 2 之间的排布顺序时，缓存设备将会得到如图 5A 和图 5B 示出的 2 个第一缓存策略组。

在一具体的实施例中，为了方便缓存设备将多个缓存策略进行排布，多个缓存策略中的每个缓存策略除包括策略本身的算法描述外，还包括策略属性的描述，简单地说，就是利用算法描述和策略属性描述来共同描述一个缓存策略。具体地，上述多个缓存策略中的每个缓存策略对应一个策略属性集合，策略属性集合包括缓存策略的类型，缓存策略的类型具体可以是步骤 S101 中提到的过滤类型、预取类型、替换类型、牺牲缓存类型、专有类型中的任一种。那么，缓存设备可以根据多个缓存策略中每个缓存策略的类型，将多个缓存策略按照预设顺序进行排布，从而得到第一缓存策略组。此处需要说明的一点是，本申请中一个缓存策略仅对应一个策略类型，以便于缓存设备可以根据缓存策略的类型将多个缓存策略进行排布。

可选的，第一缓存策略组还可以是这样得到的：缓存设备获取上述多个缓存策略之后，将这些缓存策略存储到某个预设的文件中，然后将该文件命名为第一缓存策略组。

本申请中，考虑到不同缓存策略之间可能存在不兼容的问题，比如说，过滤策略 1 用于过滤数据 A，而预取策略 4 用于将数据 A 放入缓存，那么，过滤策略 1 与预取策略 4 之间不兼容，这意味着过滤策略 1 与预取策略 4 相互冲突，如果同时运用过滤策略 1 与预取策略 4，会造成过滤策略 1 或者预取策略 4 失效，从而导致第一缓存策略组的部分功能失效。为了避免这种情况，在一具体的实施例中，缓存设备将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据之前，缓存设备还需确定第一缓存策略组的合法性，即检查多个缓存策略之间是否存在不兼容问题。

在一具体的实施例中，缓存设备根据确定第一缓存策略组的合法性，包括：缓存设备根据多个缓存策略对应的多个策略属性集合，确定第一缓存策略组的合法性；其中，每个缓存策略的策略属性集合还包括第一属性、第二属性中的至少一个，第一属性用于确定第一缓存策略组中是否存在与该缓存策略冲突的缓存策略，第二属性用于确定第一缓存策略组中能否包括多个该缓存策略。

需要说明的，由于不同的替换策略之间通常都会存在不兼容的情况，如果将不兼容的替换策略组合在一起，会导致其中的某一个替换策略失效，比如说，LRU 用于淘汰最近被访问最少的数据，LFU 用于淘汰最近被访问最不频繁的数据，如果第一缓存策略组同时包括 LRU 和 LFU，假设缓存中的某一数据为最近 20 分钟被访问次数最多数据，但该数据也是最近 2 个小时被访问最不频繁的数据，在这种情况下，依据 LRU 不应将该数据淘汰，但依据 LFU 却要将该数据淘汰。因此，替换策略的第一属性可以设置为与其他替换策略不兼容，第二属性可以设置为一个，以表示第一缓存策略组仅能包括一个替换策略。而牺牲缓存策略主要作用于被替换策略淘汰的数据，因此，当替换策略为一个时，牺牲缓存策略一般也只需要一个。因此，牺牲缓存策略的第一属性可以设置为与其他牺牲缓存策略不兼容，第二属性可以设置为一个。还需说明的，第一缓存策略组中可以包括多个相同的缓存策略，例如，假设实体先

按照数据库的行号预取数据，再按照数据库的列号预取数据，再按照数据库的行号预取数据，那么，可以依次执行预取策略 1、预取策略 2 和预取策略 1 来实现上述过程，在这种情况下，第一缓存策略组就需要 2 个预取策略 1，预取策略 1 的第二属性可以设置为多个。

在一具体的实施例中，当第一缓存策略组中的多个缓存策略是按照预设顺序排布的，上述每个缓存策略对应的策略属性集合可以具体包括前置条件、后置条件、第一属性、第二属性中的至少一个。其中，一个缓存策略的前置条件是指排布在该缓存策略前的缓存策略应满足的条件，和/或，排布在该缓存策略前的缓存策略不应满足的条件，以使得缓存设备可以根据前置条件判断第一缓存策略组中该缓存策略与其前面的缓存策略是否冲突。后置条件是指排布在该缓存策略后的缓存策略应满足的条件，和/或，排布在该缓存策略后的缓存策略不应满足的条件，以使得缓存设备可以根据后置条件判断第一缓存策略组中该缓存策略与其后面的缓存策略是否冲突。以图 5A 所示的第一缓存策略组为例，假设过滤策略 1 的后置条件为过滤数据 A，或预取策略 4 的前置条件为将数据 A 放入缓存，在这种情况下，缓存设备确定第一缓存策略组不合法。

可选的，缓存设备还可以根据前置条件、后置条件、第一属性或第二属性中的至少一个来，从多个类型的缓存策略库中选取多个缓存策略。以上述例子中替换策略的第一属性为例，缓存设备根据替换策略的第一属性，会只选择一个替换策略。又例如，假设预取策略 4 的前置条件为预取策略 4 应与过滤策略 2 同时使用，那么，当缓存设备选取了预取策略 4 之后，便会选择过滤策略 2。

在一具体的实施例中，缓存设备将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据，包括：缓存设备根据第一实体产生的数据的访问记录对第一缓存策略组进行优化，得到第二缓存策略组；然后，将第二缓存策略组运用于第一实体产生的数据。可以理解的，通过上述步骤可以得到更适应于满足第一实体的数据访问模式的第二缓存策略组，从而提高缓存的有效性。

在一具体的实施例中，缓存设备根据第一实体产生的数据的访问记录对第一缓存策略组进行优化，得到第二缓存策略组，包括：在第一缓存策略组合法的情况下，缓存设备根据第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对所述第一缓存策略组中的每个缓存策略进行迭代优化，从而得到所述第二缓存策略组。该步骤的具体内容将通过后文的步骤 S1021-S1025 进行详细叙述。可以理解的，利用启发式算法或机器学习算法来优化第一缓存策略组中的每个缓存策略，可以提高优化速度，从而使得缓存设备可以更快的获得第二缓存策略组，并运用于第一实体产生的数据。

在一具体的实施例中，缓存设备还可以将包括多个缓存策略的第一缓存策略组运用于第二实体产生的数据，其中，第一实体产生的数据与第二实体产生的数据不同。该步骤的具体过程包括：缓存设备根据第二实体产生的数据的访问记录对第一缓存策略组进行优化，得到第三缓存策略组；然后，将第三缓存策略组运用于所述第二实体产生的数据。可以理解的，缓存设备将第一缓存策略组运用于第二实体产生的数据的具体过程与将第一缓存策略组运用于第一实体产生的数据的具体过程类似，为了简便，此处不再展开详细赘述。可以看出，本申请提供的第一缓存策略组能够适用于不同实体产生的数据，也就是说，第一缓存策略组具有良好的适应能力。

下面将结合步骤 S1021-S1025 对前述步骤 S102 中缓存设备优化第一缓存策略组，从而得到第二缓存策略组的具体过程进行进一步地描述。

S1021：采集第一实体产生的数据的访问记录。

S1022: 对第一实体产生的数据的访问记录进行预处理,从而去除第一实体产生的数据中的异常数据。

在一具体的实施例中,预处理的方式包括过滤、清洗等,异常数据包括不完整数据(包括截断数据、删失数据、缺失数据等)、时间戳错误的数,地址超过地址区间的数等。

S1023: 根据预处理后的数据的访问记录,分析并评估第一缓存策略组的缓存效果。

具体实现中,将预处理后的数据输入到第一缓存策略组,经过第一缓存策略组的处理后,得到当前缓存中存储的数据,然后,再根据预处理后的数据的访问记录,确定当前缓存对应的缓存指标,从而进一步确定第一缓存策略组的缓存效果。其中,缓存指标是指衡量当前缓存的有效性的指标。可以理解的,缓存越有效,说明对应的缓存策略的缓存效果越好,因此,可以根据当前缓存对应的缓存指标来确定第一缓存策略组的缓存效果。可选的,缓存指标包括缓存命中率、缓存迁移量、读放大倍数中的至少一个,其中,缓存命中率、缓存迁移量、读放大倍数的定义可参见前述内容中的相关概念的介绍。

以缓存命中率为例,根据预处理后的数据的访问记录,确定当前缓存对应的缓存指标的具体过程为:首先,确定当前缓存中存储了预处理后的数据中的哪些数据,从而得到缓存命中的数量以及缓存没有命中的数量,然后,根据缓存命中的数量和缓存没有命中的数量计算得到当前缓存的缓存命中率。

S1024: 根据第一缓存策略组的缓存效果,确定是否需要第一缓存策略组进行优化。

在一具体的实施例中,当缓存指标不满足预设指标时,缓存设备确定需要对第一缓存策略组进行优化。当缓存指标满足预设指标时,缓存设备可以不对第一缓存策略组进行优化。其中,缓存指标满足预设指标包括以下一个或多个:缓存命中率大于预设命中率,缓存迁移量小于预设迁移量、读放大倍数小于预设倍数。预设命中率、预设迁移量以及预设倍数可以是用户设置的,也可以是缓存设备根据实际情况进行动态调整得到的,此处不作具体限定。

可以理解的,当缓存指标满足预设指标时,意味着缓存是有效的,也就是说,利用第一缓存策略组对第一实体产生的数据进行缓存时可以获得满意的缓存效果,因此为了节约资源,缓存设备可以不用对第一缓存策略组进行优化。相反的,当缓存指标不满足预设指标时,意味着缓存不具备有效性,此时需要对第一缓存策略组进行优化,从而提高缓存的有效性。

S1025、在第一缓存策略组需要进行优化的情况下,根据预处理后的数据的访问记录,对第一缓存策略组进行优化,从而得到第二缓存策略组。

在一更具体的实施例中,考虑到第一缓存策略组包括多个缓存策略,因此,第一缓存策略组的优化过程属于多目标优化问题,又由于多目标优化问题会涉及到各个子目标(此处为各个缓存策略)之间的相互制约,也就是说,一个子目标的优化可能会引起其他子目标的性能降低,因此,本申请利用启发式算法(heuristic algorithm)(例如,进化算法)或机器学习算法(例如,强化学习算法)来优化第一缓存策略组,以使得第一缓存策略组中的各个缓存策略都尽可能达到最优化。

下面遗传算法以为例,对第一缓存策略组的优化过程进行叙述。

遗传算法(genetic algorithm, GA)是一种具有高鲁棒性和广泛适应性的进化算法,通过模拟自然选择和遗传中发生的复制、交叉和变异等现象,使群体可以进化到搜索空间中越来越好的区域,从而可以产生一群最适应应用场景的个体。

假设,第一缓存策略组包括K个缓存策略,K为正整数,缓存设备利用遗传算法优化第一缓存策略组的具体过程为:首先,将第一缓存策略组看作一个群体,将第一缓存策略组中的每个缓存策略都看作一个待优化的个体,并对每个待优化的个体进行编码(即对缓存策略

的参数进行编码), 然后, 随机生成 K 个串结构的数据, 每个串结构的数据表示一个待优化的个体, 从而得到初始群体数据, 接下来, 以初始化群体数据为搜索点, 计算每个待优化的个体的适应度, 并将当前群体中适应度大于预设适应度的个体遗传到下一代群体, 然后, 通过交叉运算和变异算法产生新个体, 从而得到新一代群体, 然后确定新一代群体对应的缓存策略组的缓存指标, 当新一代群体对应的缓存策略组的缓存指标大于预设阈值时, 迭代执行上述优化过程, 直到优化后的缓存策略组的缓存指标小于或等于预设阈值时, 停止迭代, 从而得到第二缓存策略组。

下面, 通过一个具体的例子来进一步说明本申请提供的第一缓存策略组以及利用第一数据缓存方法, 具体包括如下步骤, 如图 6 所示:

步骤①: 对当前应用于实体产生的数据的缓存策略的缓存效果进行评估。

具体地, 监控当前缓存对应的缓存指标, 如果当前缓存对应的缓存指标大于预设阈值, 说明当前的缓存策略的缓存效果较好, 接下来仍可以使用当前的缓存策略。如果当前缓存对应的缓存指标小于或等于预设阈值, 意味着当前的缓存中存储了大量不符合缓存需求的数据, 也就说明当前的缓存策略的缓存效果较差, 接下来需要执行步骤②-步骤⑥来构建能够满足实体缓存需求的缓存策略。

步骤②: 采集实体产生的数据的访问日志。

步骤③: 为实体构建对应的缓存策略组。

步骤④: 对步骤③构建出的缓存策略组进行合法性检测, 并在缓存策略组合法的情况下, 对其进行缓存效果的评估。

步骤⑤: 当缓存策略组的缓存效果没有达到预设效果(即缓存指标不满足预设指标)时, 根据实体产生的数据的访问日志, 对缓存策略组进行迭代优化, 直至得到一个缓存效果达到预设效果的新的缓存策略组。

步骤⑥: 将步骤⑤得到新的缓存策略组应用到存储系统, 以用于处理实体产生的数据, 从而达到更好的缓存效果, 并进一步提高实体获取数据的速度。

可选的, 还可以省略步骤 1, 通过定期执行步骤 2-步骤 6 来优化实体对应的缓存策略, 使得缓存策略能够更加适应实体的缓存需求。

应理解, 为了简便, 上述例子并没有对每个步骤的具体实现方式进行详细叙述, 具体可参见上述方法实施例, 此处不再展开赘述。

前述内容详细阐述了本申请的方法, 为了更好地实施本申请提供的方法, 接下来将介绍本申请提供的用于配合实施上述方法的相关装置和设备。

如图 7 所示, 图 7 示出了本申请提供的一种数据缓存装置的结构示意图, 该装置 100 包括获取单元 110、缓存单元 120。

获取单元 110, 用于从多个类型的缓存策略库中获取多个缓存策略, 每个类型的缓存策略库包括的至少一个缓存策略的类型相同, 其中, 缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个, 过滤类型的缓存策略用于过滤数据, 预取类型的缓存策略用于预取数据, 替换类型的缓存策略用于从缓存中淘汰数据, 牺牲缓存类型的缓存策略用于处理从缓存中淘汰的数据。

缓存单元 120, 用于将包括多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

在一具体的实施例中, 缓存单元 120 具体用于: 根据第一实体产生的数据的访问记录对

第一缓存策略组进行优化，得到第二缓存策略组；将第二缓存策略组运用于第一实体产生的数据。

在一具体的实施例中，缓存策略库的类型还包括专有类型，专有类型的缓存策略为用户设置的缓存策略。

在一具体的实施例中，缓存单元 120 还用于：根据第二实体产生的数据的访问记录对第一缓存策略组进行优化，得到第三缓存策略组；将第三缓存策略组运用于第二实体产生的数据。

在一具体的实施例中，第一缓存策略组中的多个缓存策略是按照预设顺序排布的，当多个缓存策略包括专有类型的缓存策略时，专有类型的缓存策略的位置由用户设置。

在一具体的实施例中，数据缓存装置 100 还包括确定单元 130，确定单元 130 用于：确定第一缓存策略组的合法性。

在一具体的实施例中，多个缓存策略中的每个缓存策略对应一个策略属性集合，确定单元 130 具体用于：根据多个缓存策略对应的多个策略属性集合，确定第一缓存策略组的合法性；其中，一个缓存策略对应的策略属性集合包括第一属性、第二属性中的至少一个，第一属性用于确定第一缓存策略组中是否存在与上述一个缓存策略冲突的缓存策略，第二属性用于确定第一缓存策略组中能否包括多个上述一个缓存策略。

在一具体的实施例中，缓存单元 120 具体用于：在第一缓存策略组合法的情况下，根据第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对第一缓存策略组中的每个缓存策略进行迭代优化，从而得到第二缓存策略组。

本申请实施例的数据缓存装置 100 仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将数据缓存装置 100 的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的数据缓存装置 100 与上述方法实施例中的缓存设备属于同一构思，其具体实现过程详见上述方法实施例，例如，获取单元 110 从多个类型的缓存策略库中获取多个缓存策略的过程具体请参见上述步骤 S101，缓存单元 120 优化第一缓存策略组的过程具体请参见上述步骤 S1021-S1025，这里不再赘述。

如图 8 所示，图 8 示出了本申请提供的一种缓存设备的结构示意图，缓存设备 200 包括处理器 210、通信接口 220 和存储器 230。其中，处理器 210、通信接口 220 以及存储器 230 通过总线 240 进行耦合。

处理器 210 可以是中央处理器(central processing unit, CPU)、通用处理器、DSP、专用集成电路(application specific integrated circuit, ASIC)、现场可编程门阵列(field programmable gate array, FPGA)或者其他可编程逻辑器件(programmable logic device, PLD)、CPLD、晶体管逻辑器件、硬件部件或者其任意组合。处理器 210 可以实现或执行结合上述方法实施例中所描述的各种示例性的方法。具体的，处理器 210 读取存储器 230 中存储的程序代码，并与通信接口 220 配合执行 S101-S102、S1021-S1025 以及步骤①-⑥的部分或者全部步骤。

通信接口 220 可以为有线接口或无线接口，用于与其他模块或设备进行通信，有线接口可以是以太接口、控制器局域网络接口、局域互联网络(local interconnect network, LIN)以及 FlexRay 接口，无线接口可以是蜂窝网络接口或使用无线局域网接口等。具体的，通信接口 220 可以与其他设备连接，例如，通信接口 220 可以与存储系统相连，当处理器 210 得到第一缓存策略组之后，通过通信接口 220 可以将第一缓存策略发送给存储系统，以用于处理第

一实体产生的数据。

存储器 230 可以包括易失性存储器, 例如随机存取存储器(random access memory, RAM); 存储器 230 也可以包括非易失性存储器, 例如只读存储器(read only memory, ROM)、快闪存储器、硬盘(hard disk drive, HDD)或固态硬盘(solid state drive, SSD), 存储器 230 还可以包括上述种类的存储器的组合。存储器 230 可以存储有程序代码以及程序数据。其中, 程序代码由图 7 示出的数据缓存装置 100 中的部分或者全部单元的代码组成, 例如, 获取单元 110 的代码、缓存单元 120 的代码以及确定单元 130。程序数据由图 7 示出的数据缓存装置 100 在运行程序的过程中产生的数据, 例如, 第一实体产生的数据、缓存策略等。

总线 240 可以是控制器局域网(controller area network, CAN)或其他实现内部总线。总线 240 可以分为地址总线、数据总线、控制总线等。为了便于表示, 图 8 中仅用一条粗线表示, 但并不表示仅有一根总线或一种类型的总线。

本申请实施例中的缓存设备 200 用于执行上述方法实施例中的缓存设备执行的方法, 与上述方法实施例属于同一构思, 其具体实现过程详见上述方法实施例, 这里不再赘述。

本申请还提供了一种计算机可读存储介质, 计算机可读存储介质存储有计算机指令, 当计算机指令在计算设备(例如, 图 7 示出的数据缓存装置 100 或图 8 示出的缓存设备 200)上运行时, 使得计算设备执行上述方法实施例中的缓存设备执行的方法。

本申请还提供了一种计算机程序产品, 包括计算机程序, 当计算机程序被计算设备(例如, 图 7 示出的数据缓存装置 100 或图 8 示出的缓存设备 200)读取并执行时, 用于实现上述方法实施例中的缓存设备执行的方法。

在上述实施例中, 可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时, 可以全部或部分地以计算机程序产品的形式实现。上述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行上述计算机程序指令时, 全部或部分地产生按照本申请实施例所述的流程或功能。上述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。上述计算机指令可以存储在计算机可读存储介质中, 或者从一个计算机可读存储介质向另一个计算机可读存储介质传输, 例如, 上述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(如, 同轴电缆、光纤、数字用户线)或无线(如, 红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。上述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。上述可用介质可以是磁性介质, (如, 软盘、存储盘、磁带)、光介质(如, DVD)、或者半导体介质(如, SSD)等。在上述实施例中, 对各个实施例的描述都各有侧重, 某个实施例中未详述的部分, 可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的装置, 也可以通过其它的方式实现。例如以上所描述的装置实施例仅是示意性的, 例如所述单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 例如多个单元或组件可结合或者可以集成到另一个系统, 或一些特征可以忽略或不执行。另一点, 所显示或讨论的相互之间的间接耦合或者直接耦合或通信连接可以是通过一些接口, 装置或单元的间接耦合或通信连接, 可以是电性或其它的形式。

上述作为分离部件说明的单元可以是或者也可以不是物理上分开的, 作为单元显示的部件可以是或者也可以不是物理单元, 即可以位于一个地方, 或者, 也可以分布到多个网络单

元上。可以根据实际的需要选择其中的部分或者全部单元来实现本申请实施例的方案的目的。

另外，在本申请各实施例中的各功能单元可集成在一个处理单元中，也可以是各单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

上述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可为个人计算机、服务器或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质例如可包括：U 盘、移动硬盘、只读存储器、随机存取存储器、磁碟或光盘等各种可存储程序代码的介质。

权 利 要 求 书

1. 一种数据缓存方法，其特征在于，所述方法包括：

从多个类型的缓存策略库中获取多个缓存策略，每个类型的缓存策略库包括的至少一个缓存策略的类型相同，其中，所述缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个，所述过滤类型的缓存策略用于过滤数据，所述预取类型的缓存策略用于预取数据，所述替换类型的缓存策略用于从缓存中淘汰数据，所述牺牲缓存类型的缓存策略用于处理从所述缓存中淘汰的数据；

将包括所述多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

2. 根据权利要求1所述的方法，其特征在于，所述将包括所述多个缓存策略的第一缓存策略组运用于第一实体产生的数据，包括：

根据所述第一实体产生的数据的访问记录对所述第一缓存策略组进行优化，得到第二缓存策略组；

将所述第二缓存策略组运用于所述第一实体产生的数据。

3. 根据权利要求1或2所述的方法，其特征在于，所述缓存策略库的类型还包括专有类型，所述专有类型的缓存策略为用户设置的缓存策略。

4. 根据权利要求1-3任一项所述的方法，其特征在于，所述方法还包括：

根据第二实体产生的数据的访问记录对所述第一缓存策略组进行优化，得到第三缓存策略组；

将所述第三缓存策略组运用于所述第二实体产生的数据。

5. 根据权利要求3所述的方法，其特征在于，所述第一缓存策略组中的多个缓存策略是按照预设顺序排布的，当所述多个缓存策略包括所述专有类型的缓存策略时，所述专有类型的缓存策略的位置由所述用户设置。

6. 根据权利要求1-5任一项所述的方法，其特征在于，在所述将包括所述多个缓存策略的第一缓存策略组运用于第一实体产生的数据之前，所述方法还包括：

确定所述第一缓存策略组的合法性。

7. 根据权利要求6所述的方法，其特征在于，所述多个缓存策略中的每个缓存策略对应一个策略属性集合，所述确定所述第一缓存策略组的合法性，包括：

根据所述多个缓存策略对应的多个策略属性集合，确定所述第一缓存策略组的合法性；

其中，一个缓存策略对应的策略属性集合包括第一属性、第二属性中的至少一个，所述第一属性用于确定所述第一缓存策略组中是否存在与所述一个缓存策略冲突的缓存策略，所述第二属性用于确定所述第一缓存策略组中能否包括多个所述一个缓存策略。

8. 根据权利要求6或7所述的方法，其特征在于，所述根据所述第一实体产生的数据的访问记录对所述第一缓存策略组进行优化，得到第二缓存策略组，包括：

在所述第一缓存策略组合法的情况下，根据所述第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对所述第一缓存策略组中的每个缓存策略进行迭代优化，从而得到所述第二缓存策略组。

9. 一种数据缓存装置，其特征在于，所述装置包括：

获取单元，用于从多个类型的缓存策略库中获取多个缓存策略，每个类型的缓存策略库包括的至少一个缓存策略的类型相同，其中，所述缓存策略库的类型包括过滤类型、预取类型、替换类型、牺牲缓存类型中的至少一个，所述过滤类型的缓存策略用于过滤数据，所述预取类型的缓存策略用于预取数据，所述替换类型的缓存策略用于从缓存中淘汰数据，所述

牺牲缓存类型的缓存策略用于处理从所述缓存中淘汰的数据；

缓存单元，用于将包括所述多个缓存策略的第一缓存策略组运用于第一实体产生的数据。

10. 根据权利要求9所述的装置，其特征在于，所述缓存单元具体用于：

根据所述第一实体产生的数据的访问记录对所述第一缓存策略组进行优化，得到第二缓存策略组；

将所述第二缓存策略组运用于所述第一实体产生的数据。

11. 根据权利要求9或10所述的装置，其特征在于，所述缓存策略库的类型还包括专有类型，所述专有类型的缓存策略为用户设置的缓存策略。

12. 根据权利要求9-11任一项所述的装置，其特征在于，所述缓存单元还用于：

根据第二实体产生的数据的访问记录对所述第一缓存策略组进行优化，得到第三缓存策略组；

将所述第三缓存策略组运用于所述第二实体产生的数据。

13. 根据权利要求11所述的装置，其特征在于，所述第一缓存策略组中的多个缓存策略是按照预设顺序排布的，当所述多个缓存策略包括所述专有类型的缓存策略时，所述专有类型的缓存策略的位置由所述用户设置。

14. 根据权利要求9-13任一项所述的装置，其特征在于，所述装置还包括确定单元，所述确定单元用于：确定所述第一缓存策略组的合法性。

15. 根据权利要求14所述的装置，其特征在于，所述多个缓存策略中的每个缓存策略对应一个策略属性集合，所述确定单元具体用于：

根据所述多个缓存策略对应的多个策略属性集合，确定所述第一缓存策略组的合法性；

其中，一个缓存策略对应的策略属性集合包括第一属性、第二属性中的至少一个，所述第一属性用于确定所述第一缓存策略组中是否存在与所述一个缓存策略冲突的缓存策略，所述第二属性用于确定所述第一缓存策略组中能否包括多个所述一个缓存策略。

16. 根据权利要求14或15所述的装置，其特征在于，所述缓存单元具体用于：

在所述第一缓存策略组合法的情况下，根据所述第一实体产生的数据的访问记录，利用启发式算法或机器学习算法对所述第一缓存策略组中的每个缓存策略进行迭代优化，从而得到所述第二缓存策略组。

17. 一种缓存设备，其特征在于，所述缓存设备包括处理器和存储器，所述处理器执行所述存储器中的代码以实现权利要求1至8任一项所述的方法。

18. 一种计算机可读存储介质，其特征在于，存储有计算机指令，所述计算机指令用于实现权利要求1至8任一项所述的方法。

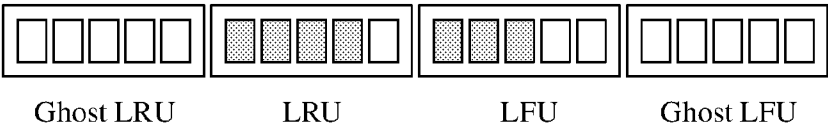


图 1A

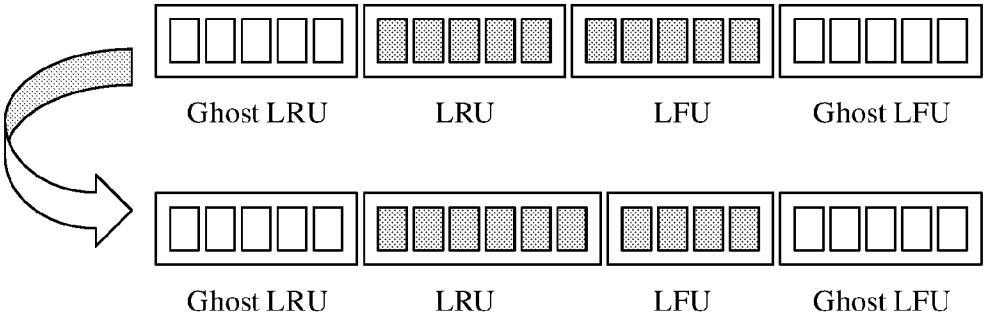


图 1B

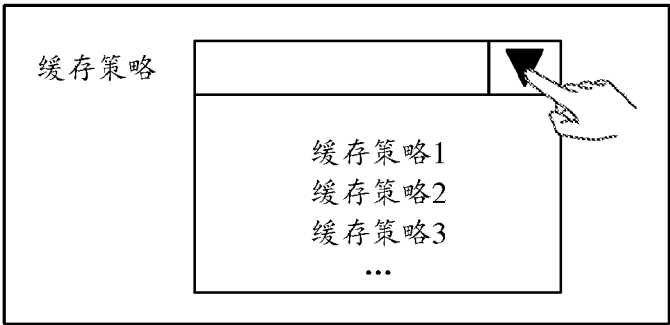


图 2A

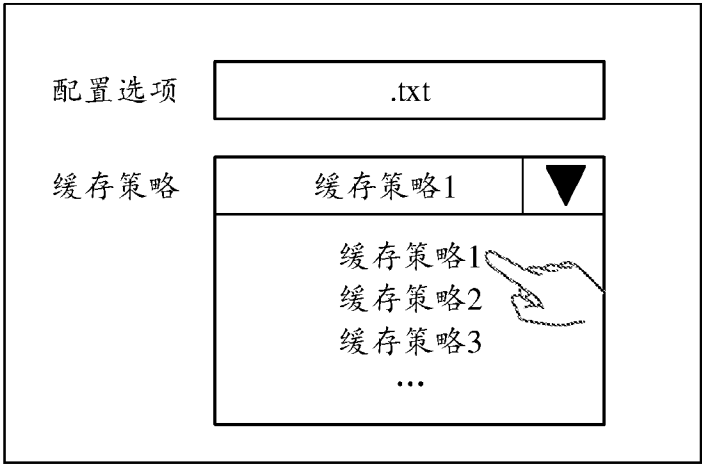


图 2B

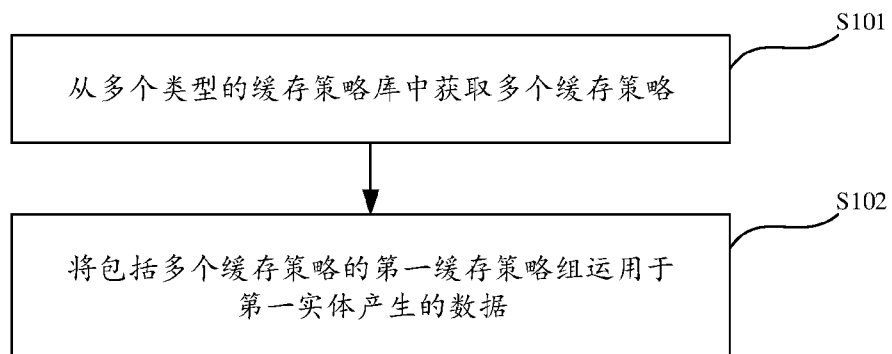


图 3

过滤策略库

☐ 过滤策略1

☐ 过滤策略2

☐ 过滤策略3

☐ 无

预取策略库

☐ 预取策略1

☐ 预取策略2

☐ 预取策略3

☐ 预取策略4

☐ 无

替换策略库

☐ 替换策略1

☐ 替换策略2

☐ 替换策略3

☐ 替换策略4

☐ 替换策略5

☐ 无

牺牲缓存策略库

☐ 牺牲缓存策略1

☐ 牺牲缓存策略2

☐ 无

专有策略库

☐ 专有策略1

☐ 专有策略2

☐ 无

图 4A

过滤策略库

☐ 过滤策略1

☐ 过滤策略2

☐ 过滤策略3

☐ 无

预取策略库

☐ 预取策略1

☐ 预取策略2

☐ 预取策略3

☐ 预取策略4

☐ 无

替换策略库

☐ 替换策略1

☐ 替换策略2

☐ 替换策略3

☐ 替换策略4

☐ 替换策略5

☐ 无

牺牲缓存策略库

☐ 牺牲缓存策略1

☐ 牺牲缓存策略2

☐ 无

专有策略库

☐ 专有策略1

☐ 专有策略2

☐ 无

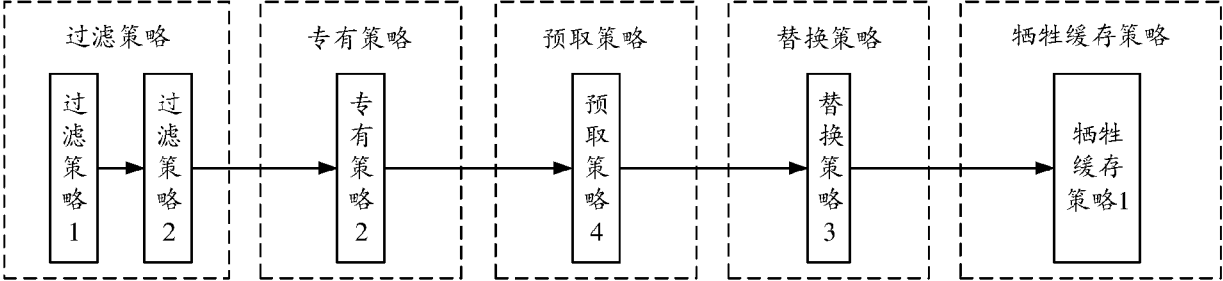
配置选项

▼

缓存内容
缓存时间
缓存优先级
...

目录A

图 4B



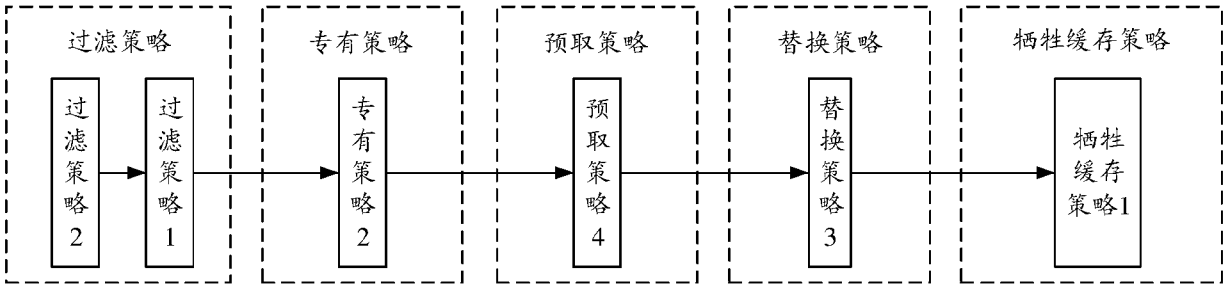


图 5B

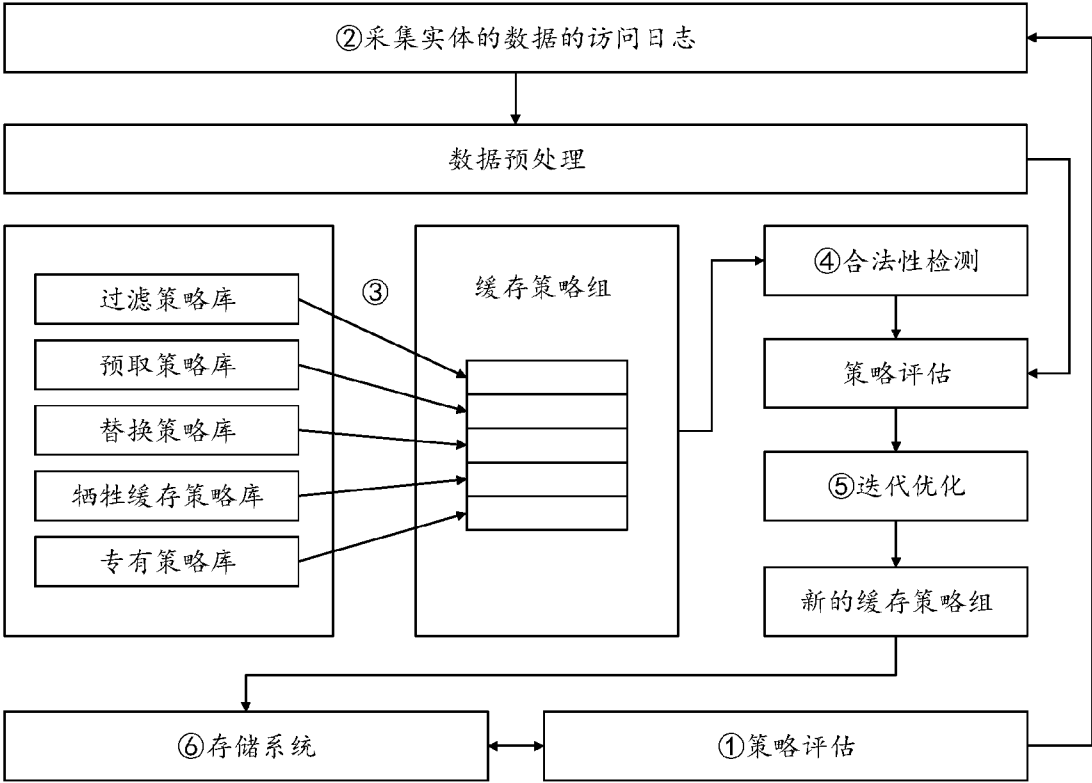


图 6

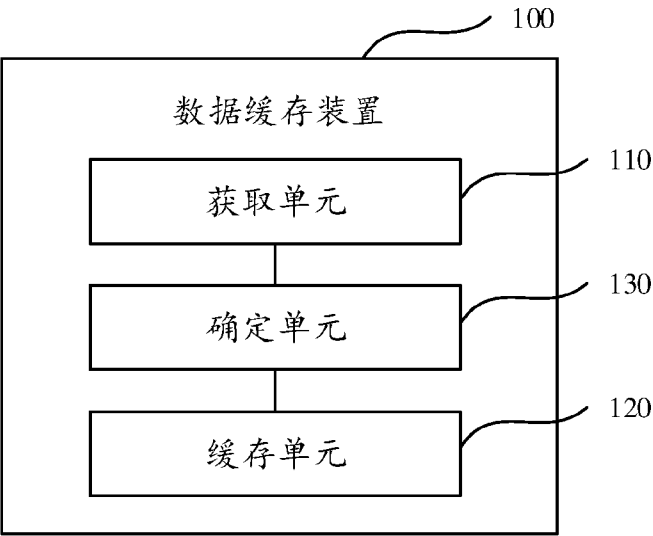


图 7

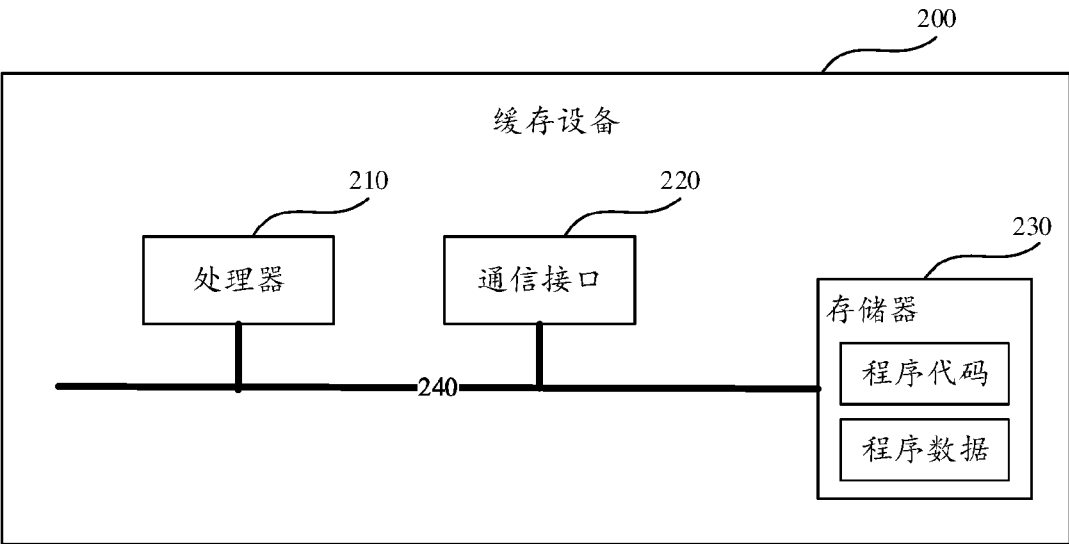


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/071079

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/2455(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, WPI, EPODOC, CNPAT, IEEE: 缓存, 策略, 规则, 多种, 库, 数据库, 单一, 过滤, 牺牲, 替换, 预取, cache, rule, strategy, multiple, database, filter, replace, prefetch, sacrifice

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106708636 A (BEIJING QIHOO TECHNOLOGY CO., LTD.) 24 May 2017 (2017-05-24) description paragraphs [0040]-[0047], [0091], claim 1	1-18
A	US 2019124174 A1 (ALIBABA GROUP HOLDING LIMITED) 25 April 2019 (2019-04-25) entire document	1-18
A	CN 110929195 A (VIEWHIGH KANGXIN (BEIJING) TECHNOLOGY CO., LTD.) 27 March 2020 (2020-03-27) entire document	1-18
A	CN 106921713 A (CHINA MOBILE GROUP SHANGHAI CO., LTD.) 04 July 2017 (2017-07-04) entire document	1-18



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

29 March 2022

Date of mailing of the international search report

06 April 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/071079

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	106708636	A	24 May 2017	None	
US	2019124174	A1	25 April 2019	CN 106997351 A	01 August 2017
				WO 2017124972 A1	27 July 2017
CN	110929195	A	27 March 2020	None	
CN	106921713	A	04 July 2017	None	

国际检索报告

国际申请号

PCT/CN2022/071079

A. 主题的分类

G06F 16/2455 (2019. 01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNKI, WPI, EPDOC, CNPAT, IEEE:缓存, 策略, 规则, 多种, 库, 数据库, 单一, 过滤, 牺牲, 替换, 预取, cache, rule, strategy, multiple, database, filter, replace, prefetch, sacrifice

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106708636 A (北京奇虎科技有限公司) 2017年5月24日 (2017 - 05 - 24) 说明书第[0040]-[0047], [0091]段, 权利要求1	1-18
A	US 2019124174 A1 (ALIBABA GROUP HOLDING LIMITED) 2019年4月25日 (2019 - 04 - 25) 全文	1-18
A	CN 110929195 A (望海康信北京科技股份有限公司) 2020年3月27日 (2020 - 03 - 27) 全文	1-18
A	CN 106921713 A (中国移动通信集团上海有限公司) 2017年7月4日 (2017 - 07 - 04) 全文	1-18

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年3月29日

国际检索报告邮寄日期

2022年4月6日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

董刚

电话号码 86-(10)-53961348

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/071079

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106708636	A	2017年5月24日	无	
US	2019124174	A1	2019年4月25日	CN	106997351 A 2017年8月1日
				WO	2017124972 A1 2017年7月27日
CN	110929195	A	2020年3月27日	无	
CN	106921713	A	2017年7月4日	无	