

(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
2021 年 3 月 4 日 (04.03.2021)



(10) 国际公布号
WO 2021/036452 A1

- (51) 国际专利分类号:
G06F 16/2455 (2019.01)
- (21) 国际申请号: PCT/CN2020/097839
- (22) 国际申请日: 2020 年 6 月 24 日 (24.06.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910795939.4 2019年8月27日 (27.08.2019) CN
- (71) 申请人: 苏宁云计算有限公司 (SUNING CLOUD COMPUTING CO., LIMITED) [CN/CN]; 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu 210042 (CN)。
- (72) 发明人: 汪凯(WANG, Kai); 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu 210042 (CN)。 张盼盼(ZHANG, Panpan); 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu

210042 (CN)。 韩振旭(HAN, Zhenxu); 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu 210042 (CN)。 李成(LI, Cheng); 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu 210042 (CN)。 孙迁(SUN, Qian); 中国江苏省南京市玄武区徐庄软件园苏宁大道 1-1 号, Jiangsu 210042 (CN)。

(74) 代理人: 北京律和信知识产权代理事务所 (普通合伙) (BEIJING LAWSING IP FIRM); 中国北京市海淀区海淀大街 38 号银科大厦 802 室, Beijing 100080 (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, 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,

(54) Title: REAL-TIME DATA DEDUPLICATION COUNTING METHOD AND DEVICE

(54) 发明名称: 实时数据去重计数方法及装置

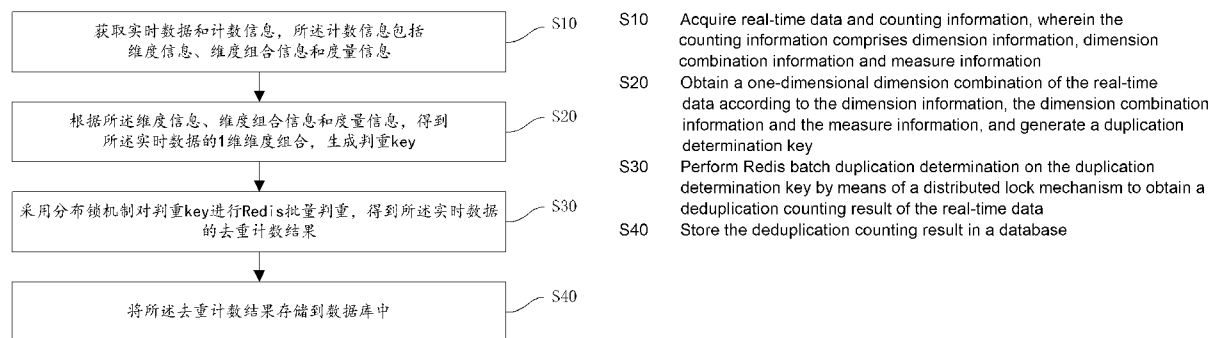


图2

(57) Abstract: A real-time data deduplication counting method and device. The accurate calculation of a measure deduplication counting value of real-time data can be realized, and a small memory space is occupied in a deduplication process. The method comprises: acquiring real-time data and counting information, wherein the counting information comprises dimension information, dimension combination information and measure information (S10); obtaining a one-dimensional dimension combination of the real-time data according to the dimension information, the dimension combination information and the measure information, and generating a duplication determination key (S20); performing Redis batch duplication determination on the duplication determination key by means of a distributed lock mechanism to obtain a deduplication counting result of the real-time data (S30); and storing the deduplication counting result in a database (S40). According to the method, the uniqueness of a measure field of real-time data under each dimension combination can be guaranteed, and an accurate deduplication counting value of the real-time data can be obtained through duplication determination. Only the duplication determination key of a one-dimensional dimension combination needs to be stored for duplication determination, a large number of memory resources are released, the length of the duplication determination key is fixed, and the Redis memory usage amount can be estimated.



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,
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))。

(57) 摘要: 一种实时数据去重计数方法及装置, 可实现实时数据的度量去重计数值的精准计算, 同时去重过程占用内存空间小。该方法包括: 获取实时数据和计数信息, 所述计数信息包括维度信息、维度组合信息和度量信息 (S10); 根据所述维度信息、维度组合信息和度量信息, 得到所述实时数据的1维维度组合, 生成判重key (S20); 采用分布锁机制对判重key进行Redis批量判重, 得到所述实时数据的去重计数结果 (S30), 将所述去重计数结果存储到数据库中 (S40)。通过该方法可以保证各维度组合下实时数据的度量字段的唯一性, 判重得到实时数据的精准去重计数值。只需保存1维维度组合的判重key进行判重, 大量内存资源得到释放, 且判重key的长度固定, 可预估Redis内存使用量。

实时数据去重计数方法及装置

技术领域

本发明属于数据分析技术领域，具体涉及一种实时数据去重计数方法及装置。

背景技术

在大数据分析中，经常需要对某度量字段进行去重统计，而目前常用的联机分析处理（OLAP）工具无法做到实时数据的去重计数值的精准计算。而且使用常规方案去重，即所有维度组合下各维度维值集合的笛卡尔积为判重 key，而一个度量字段下每种维度组合的判重 key 的最大数目，是该维度组合下各维度基数的乘积。比如一个 Cube^[1]包含 time、item、location 和 supplier 四个维度，如图 1 所示，有 16 种维度组合，即 16 个 Cuboid^[2]。若四个维度的基数分别为 a、b、c、d，则组合（time，item）的判重 Key 的最大数目是 $a*b$ ，组合（time，item，location）的判重 Key 的最大数目是 $a*b*c$ ，组合（time，item，location，supplier）的判重 Key 的最大数目是 $a*b*c*d$ 等。以此类推，判重 Key 的最大总数目为 $1+(a+b+c+d)+(a*b+a*c+a*d+b*c+b*d+c*d)+(a*b*c+a*b*d+a*c*d+b*c*d)+a*b*c*d$ 。由此可见，目前常规去重方法会占用较大的内存空间，造成严重的资源浪费，而且内存使用量估算困难。

发明内容

本发明的实施例提供一种实时数据去重计数方法及装置，以便解决的问题。

为达到上述目的，本发明的实施例采用如下技术方案：

第一方面，本发明的实施例提供一种实时数据去重计数方法，包括：

获取实时数据和计数信息，所述计数信息包括维度信息、维度组合信息和

度量信息；

根据所述维度信息、维度组合信息和度量信息，将所述实时数据的所有维度组合分解成1维维度组合，生成判重key；

采用分布锁机制对判重 key 进行 Redis 批量判重，得到所述实时数据的去重计数结果。

结合第一方面，在第一方面的第一种可能的实现方式中，采用分布锁机制对判重key进行Redis批量判重，得到所述实时数据的去重计数结果，具体包括：

判断所述实时数据中各条数据是否存在竞争，若不存在，则所有判重key进行Redis批量判重；若存在，则存在竞争的其中一条数据加锁，该条数据的判重key进行Redis批量判重，处理好后释放锁，其他数据轮询加锁，加锁成功的数据的判重key进行Redis批量判重，处理后释放锁，循环直至存在竞争的所有数据的判重key均完成Redis批量判重；其中，所述Redis批量判重具体是，Redis批量判断判重key是否存在，若不存在则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1，生成预计算结果；

将预计算结果中相同维度组合下相同度量字段值的度量值进行累加，生成所述实时数据的去重计数结果。

结合第一方面的或者第一方面的第一种可能的实现方式，在第一方面的第二种可能的实现方式中，所述方法还包括：将所述实时数据的去重计数结果存储到数据库中。

结合第一方面的第二种可能的实现方式，在第一方面的第三种可能的实现方式中，所述将所述实时数据的去重计数结果存储到数据库中，具体包括：

将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，依次查找数据库中是否存在相同的查找字段，若存在则将相同的查找字段对应

的值相加后存入数据库，若不存在则将查找字段和其对应的值存入数据库。

结合第一方面，在第一方面的第四种可能的实现方式中，所述维度信息包括维度编码，维度组合信息包括维度组合维度编码，度量信息包括度量编码。

结合第一方面的第四种可能的实现方式，在第一方面的第五种可能的实现方式中，生成判重key具体是，将1维维度组合的维度编码拼接当日日期、实时数据的度量字段值和实时数据的维度字段值作为判重key。

第二方面，本发明的实施例提供一种实时数据去重计数装置，包括：

获取模块，用于获取实时数据和计数信息，所述计数信息包括维度信息、维度组合信息和度量信息；

判重key生成模块，用于根据所述维度信息、维度组合信息和度量信息，得到所述实时数据的1维维度组合，生成判重key；

计数模块，用于采用分布锁机制对判重 key 进行 Redis 批量判重，得到所述实时数据的去重计数结果。

结合第二方面，在第二方面的第一种可能的实现方式中，所述计数模块包括：

分布计算单元，用于判断所述实时数据中各条数据是否存在竞争，若不存在，则将所有判重key发送给Redis批量判重单元；若存在，则存在竞争的其中一条数据加锁，将该条数据的判重key发送给Redis批量判重单元，处理好后释放锁，其他数据轮询加锁，将加锁成功的数据的判重key发送给Redis批量判重单元，处理后释放锁，循环直至存在竞争的所有数据的判重key均发送给Redis批量判重单元；

Redis批量判重单元，用于Redis批量判断判重key是否存在，若不存在则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1，生成预计算

结果；

累加单元，用于将 Redis 批量判重单元得到预计算结果中相同维度组合下相同度量字段值的度量值进行累加，生成所述实时数据的去重计数结果。

结合第二方面或者第二方面的第一种可能的实现方式，在第二方面的第二种可能的实现方式中，所述装置还包括：

存储模块，用于将所述实时数据的去重计数结果存储到数据库中。

结合第二方面的第二种可能的实现方式，在第二方面的第三种可能的实现方式中，所述存储模块进一步用于将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，依次查找数据库中是否存在相同的查找字段，若存在则将相同的查找字段对应的值相加后存入数据库，若不存在则将查找字段和其对应的值存入数据库。

本发明实施例提供的实时数据去重计数方法及装置，可实现实时数据的度量去重计数值的精准计算，同时去重过程占用内存空间小。本发明实施例中，将实时数据的所有维度组合分解成 1 维维度组合，根据 1 维维度组合生成判重 key，使用分布式锁的控制对判重 key 进行 Redis^[3]批量去重处理，可以保证各维度组合下实时数据的度量字段的唯一性，判重得到实时数据的精准去重计数值。同时只需保存 1 维维度组合的判重 key 进行判重，相比于现有去重方法，判重 key 的数量少，大量内存资源得到释放，且判重 key 的长度固定，容易预估 Redis 内存使用量。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例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它的附图。

图1为一个包含四个维度的Cube示例图；

图2为本发明实施例提供的一种实时数据去重计数方法的流程图；

图3为运行本发明实施例方法流程的系统架构图；

图4为线程竞争关系图；

图5为本发明另一实施例提供的一种实时数据去重计数装置的结构框图。

具体实施方式

为使本领域技术人员更好地理解本发明的技术方案，下面结合附图和具体实施方式对本发明作进一步详细描述。下文中将详细描述本发明的实施方式，所述实施方式的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语（包括技术术语和科学术语）具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语应该被理解为具有与现有技术上下文中的意义一致的意义，并且除非像这里一样定义，不会用理想化或过于正式的含义来解释。

下面结合图2具体说明本发明实施例的方法流程。

本发明实施例提供的一种实时数据去重计数方法，包括以下步骤：

S10，获取实时数据和计数信息，所述计数信息包括维度信息、维度组合信息和度量信息。

根据本发明的一个实施例，采用Kafka^[4]作为数据源，从Kafka批量获取实时数据，获得的批量实时数据中包含多条数据，每条数据至少包括日期、度量字

段值和维度字段值。

查询MySQL^[6]的配置信息，获取计数信息，计数信息包括维度信息、维度组合信息和度量信息。其中所述的维度主要是一些在统计数据中的非统计项，例如：大区，品类等字段，对应的维度组合有大区Cuboid，品类Cuboid和大区-品类Cuboid。所述的度量是统计数据中的统计项，例如会员号、访问量、订单号等字段。

在一些实施例中，为了便于在整个方法流程中记录处理，每个维度设有对应的维度编码，每个维度组合设有对应的维度组合编码，每个度量设有对应的度量编码。所述维度信息包括维度编码，所述维度组合信息包括维度组合编码。每个度量有对应的度量编码，所述度量信息包括度量编码。

S20，根据所述维度信息、维度组合信息和度量信息，得到所述实时数据的所有1维维度组合，生成判重key。

遍历所述批量实时数据中每条数据的所有Cuboid，将每个Cuboid分解成一个或多个1维Cuboid，不相同的Cuboid各生成一个判重key。例如，若选择A和B两个维度进行度量，则所述实时数据构成的Cube包含A和B两个维度，其维度组合就有4种，分别是一个0维Cuboid，两个1维Cuboid（A Cuboid和B Cuboid）和一个2维Cuboid（A-B Cuboid）。将A-B Cuboid分解成A Cuboid和B Cuboid，A Cuboid生成一个判重key，B Cuboid生成一个判重key，故每条数据均生成两个判重key。在实际处理中，只需选择A Cuboid和B Cuboid，A Cuboid生成判重key1，B Cuboid生成判重key2，与判重key1关联的有A Cuboid和可分解为A Cuboid的A-B Cuboid，与判重key2关联的有B Cuboid和可分解为B Cuboid的A-B Cuboid，只需判重key1和key2，该条数据的所有Cuboid均可得到去重。

根据本发明的一个实施例，将1维Cuboid的维度编码拼接日期、实时数据的

度量字段值和实时数据的维度字段值作为判重key。

常规的去重方法中将主题、消费组、日期、度量字段值、维度组合编码和维值拼接作为判重key，由于Cuboid是多维度组合，判重key的长度不固定，且每个Cuboid对应一组判重key，导致了Redis内存使用量的预估困难。而本发明实施例只保存1维Cuboid的判重key，判重key的长度固定，不会有因Cuboid的不同维值组合导致判重Key的膨胀问题，比常规去重方法的判重key的长度短且个数少，占用内存少，内存使用量可估算。

S30，采用分布锁机制对判重key进行Redis批量判重，得到所述实时数据去重计数结果。

根据本发明的一个实施例，判断所述实时数据中每条数据是否存在竞争，若不存在，则所有数据的判重key进行判重处理；若存在，则存在竞争的其中一条数据加锁，该条数据的判重key进行判重处理，处理好后释放锁，其他数据轮询尝试加锁，加锁成功的数据的判重key进行判重处理，处理后释放锁，循环直至存在竞争的所有数据的判重key均完成判重处理。

其中，通过以下方法判断每条实时数据是否存在竞争，具体是每条数据均生成一个锁key，将当日日期拼接度量字段值作为锁key，若每条数据的锁key均不相同，则该批量实时数据的各条数据不存在竞争，若锁key相同，则具有相同锁key的数据存在竞争。

Redis批量去重处理具体是，Redis批量判断判重key是否存在，若不存在则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1。

将所述实时数据的相同维度组合下相同度量字段值的度量值进行累加，得到所述实时数据的各维度组合下的去重度量计数值，生成所述实时数据的去重计数结果。

通过1维Cuboid的判重key进行判重，如果具有相同度量字段值的几条数据分别由不同的线程并发处理，则会导致判重结果有误，而本发明实施例使用分布锁机制控制并发，可预防上述情况的发生，使得只需要存储1维Cuboid维度值就可以达到判重的目的，大量内存资源得到释放。

根据本发明的一个实施例，所述方法还包括：

S40，将所述去重计数结果存储到数据库中。

具体的，将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，依次查找数据库中是否存在相同的查找字段，若存在则将相同的查找字段对应的值相加存入数据库，若不存在则将查找字段和其对应的值存入数据库。

本发明实施例将实时数据的去重计数结果存入数据库，可实现实时、快速查询去重计数结果。

本发明实施例提供的方法流程具体可以运行在图3所示的系统中，采用Kafka作为数据源，HBase^[6]作为存储引擎，Spark Streaming^[7]作为计算引擎，执行本发明实施例提供的方法流程。

下面以大区 and 品类两个维度计算度量买家数为例介绍本发明实施例方法的具体流程。

101) 从Kafka批量获取实时数据，获取的批量订单数据如表1所示：

表1

订单号	订单行号	付款日期	会员号	大区	品类
19052001	1905200101	20190520	52001	10001	00001
19052002	1905200201	20190520	52002	10001	00001
19052002	1905200202	20190520	52002	10001	00002
19052003	1905200301	20190520	52003	10002	00001

以上数据是参与2019年5月20日买家数计算的实时订单数据。其中，“大区”下的字段值10001表示南京大区，字段值10002表示无锡大区，“品类”下的字段值00001表示空调，字段值00002表示冰洗。

102) 查询MySQL中的配置信息，获取维度信息：大区（编码：WD1020）和品类（编码：WD1029），维度组合信息：大区Cuboid（编码：VV254），品类Cuboid（编码：VV151）和大区-品类Cuboid（编码：VV153），以及度量信息：买家数（编码：ZB_YY_0001，SJ_01）。需要说明的是，本例是计算度量买家数，买家数对应表1中的会员号，即统计项是表1中的会员号。

201) 根据维度信息、维度组合信息和度量信息，生成上述批量实时数据中每条订单的判重Key。

具体是，表1中的订单数据具有3个Cuboid，分别为大区Cuboid（编码：VV254），品类Cuboid（编码：VV151）和大区-品类Cuboid（编码：VV153），将大区-品类Cuboid分解为大区Cuboid和品类Cuboid，大区Cuboid的维度编码拼接当日日期、度量字段值和维度字段值作为判重Key，品类Cuboid的维度编码拼接当日日期、解析出的度量字段值和维度字段值作为判重Key，故上述每条订单均生成两个判重Key，如表2所示。其中，度量字段值为表1中“会员号”下的字段值，维度字段值为表1中“大区”或/和“品类”下的字段值。

表2

订单行号	判重key
1905200101	20190520-52001-WD1020-10001
	20190520-52001-WD1029-00001
1905200201	20190520-52002-WD1020-10001
	20190520-52002-WD1029-00001

1905200202	20190520-52002-WD1020-10001
	20190520-52001-WD1029-00002
1905200301	20190520-52003-WD1020-10002
	20190520-52003-WD1029-00001

202) 生成上述批量实时数据中每条订单的锁Key。具体是，将当日日期拼接度量字段值作为锁Key，如表3所示。

表3

订单行号	锁key
1905200101	20190520-52001
1905200201	20190520-52002
1905200202	20190520-52002
1905200301	20190520-52003

203) 将上述所有判重Key批量放入Redis中，进行并发去重处理，得到预计算结果。所述预计算结果包括去重结果Key和对应的去重结果，其中将维度组合编码，当日日期拼接维度字段值作为去重结果Key，将度量编码和度量值作为去重结果。

具体是，根据表3所示的四条订单的锁Key，判断该批次实时数据的四条订单是否存在竞争。由于订单1905200101和订单1905200301的锁Key不相同，故这两条订单不存在竞争，两条订单直接进行判重处理。由于订单1905200201和订单1905200202的锁Key相同，故这两条订单存在竞争，如图4所示，订单1905200201加锁，进行判重处理，处理结束后释放锁。订单1905200202轮询尝试加锁，加锁成功后进行判重处理。

其中判重处理的具体步骤是：以订单1905200101为例，判断得到判重

Key20190520-52001-WD1020-10001不存在，即大区Cuboid的度量字段值（会员号）不重复，故大区视觉下有一个买家，对应的度量值为1，生成对应的预计算结果是（“VV254, 20190520-10001”，{“ZB_YY_0001, SJ_01”: 1}）。接着判断得到判重Key20190520-52001-WD1029-00001不存在，即品类Cuboid的度量字段值（会员号）不重复，故品类视觉下有一个买家，对应的度量值为1，生成对应的预计算结果为（“VV151, 20190520-00001”，{“ZB_YY_0001, SJ_01”: 1}）。由于大区Cuboid和品类Cuboid的度量字段值（会员号）均不重复，故大区-品类Cuboid的度量字段值（会员号）也不重复，则大区-品类视觉下有一个买家，对应的度量值为1，生成对应的预计算结果为（“VV153, 20190520-10001-00001”，{“ZB_YY_0001, SJ_01”: 1}）。以同样的方法得到订单1905200201的预计算结果和订单1905200202的预计算结果。

而订单1905200202判重处理时，判断得到判重

Key20190520-52002-WD1020-10001存在，即大区Cuboid的度量字段值（会员号）重复。接着判断得到判重Key20190520-52002-WD1029-00002不存在，即品类Cuboid的度量字段值（会员号）不重复，则品类视觉下有一个买家，对应的度量值为1，生成对应的预计算结果为（“VV151, 20190520-00002”，{“ZB_YY_0001, SJ_01”: 1}）。因为品类Cuboid的度量字段值不重复，故大区-品类Cuboid的度量字段值也不重复，大区-品类视觉有一个买家，对应的度量值为1，生成对应的预计算结果为（“VV152, 20190520-10001-00002”，{“ZB_YY_0001, SJ_01”: 1}）。

得到表1所示的实时订单数据的预计算结果如表4所示。

表4

订单行号	预计算结果
------	-------

1905200101	(“VV254, 20190520-10001”, { “ZB_YY_0001, SJ_01”: 1}) (“VV151, 20190520-00001”, { “ZB_YY_0001, SJ_01”: 1}) (“VV153, 20190520-10001-00001”, { “ZB_YY_0001, SJ_01”: 1})
1905200201	(“VV254, 20190520-10001”, { “ZB_YY_0001, SJ_01”: 1}) (“VV151, 20190520-00001”, { “ZB_YY_0001, SJ_01”: 1}) (“VV153, 20190520-10001-00001”, { “ZB_YY_0001, SJ_01”: 1})
1905200202	(“VV151, 20190520-00002”, { “ZB_YY_0001, SJ_01”: 1}) (“VV153, 20190520-10001-00002”, { “ZB_YY_0001, SJ_01”: 1})
1905200301	(“VV254, 20190520-10002”, { “ZB_YY_0001, SJ_01”: 1}) (“VV151, 20190520-00001”, { “ZB_YY_0001, SJ_01”: 1}) (“VV153, 20190520-10002-00001”, { “ZB_YY_0001, SJ_01”: 1})

204) 将上述预计算结果中, 相同的去重结果Key对应的度量值累加, 具体是, 将表4中所有“VV254, 20190520-10001”对应的度量值累加, 得到去重度量计数值是2, “VV254, 20190520-10002”的去重度量计数值是1, “VV151, 20190520-00001”的去重度量计数值是3, “VV151, 20190520-00002”的去重度量计数值是1, “VV153-20190520-10001-00001”的去重度量计数值是2, “VV153-20190520-10001-00002”的去重度量计数值是1, “VV153-20190520-10002-00001”的去重度量计数值是1。

301) 将上述累加结果存入HBase。

具体是以维度组合编码拼接度量编码、当日日期和维度字段值的形式映射成HBase的RowKey, 去重度量计数值映射成HBase的Value, 如表5所示。然后通过Rowkey批量查询HBase, 存在将Value求和入库, 不存在则直接入库。

表5

RowKey	Quota:ZB_XS_0001_SJ_01
VV253-ZB_YY_0001-20190520-10001	2

VV253-ZB_YY_0001-20190520-10002	1
VV253-ZB_YY_0001-20190520-00001	3
VV253-ZB_YY_0001-20190520-00002	1
VV253-ZB_YY_0001-20190520-10001-00001	2
VV253-ZB_YY_0001-20190520-10001-00002	1
VV253-ZB_YY_0001-20190520-10002-00001	1

本发明实施例提供的实时数据去重计数方法，可实现实时数据的度量去重计数值的精准计算，同时去重过程占用内存空间小。本发明实施例中，将实时数据的所有维度组合分解成1维维度组合，根据1维维度组合生成判重key，使用分布式锁的控制对判重key进行Redis批量去重处理，可以保证各维度组合下实时数据的度量字段的唯一性，判重得到实时数据的精准去重计数值。同时只需保存1维维度组合的判重key进行判重，相比于现有去重方法，判重key的数量少，大量内存资源得到释放，且判重key的长度固定，可预估Redis内存使用量。以图1为例，采用本发明实施例的去重计数方法，组合（time，item）的判重Key的最大数目是a+b，组合（time，item，location）的判重Key的最大数目是a+b+c，组合（time，item，location，supplier）的判重Key的最大数目是a+b+c+d等。16种维度组合只需要0维和1维Cuboid，判重Key的最大总数目为1+(a+b+c+d)，即各维度基数总和加1。

本发明另一个实施例提供了一种实时数据去重计数装置，如图5所示，包括：

获取模块，用于获取实时数据、维度信息、维度组合信息和度量信息；

判重key生成模块，用于根据所述维度信息、维度组合信息和度量信息，得到所述实时数据的1维维度组合，生成判重key；

计数模块，用于采用分布锁机制对判重key进行Redis批量判重，得到所述

实时数据的去重计数结果。

可选的，所述计数模块包括：

分布计算单元，用于判断所述实时数据中各条数据是否存在竞争，若不存在，则将所有判重key发送给Redis批量判重单元；若存在，则存在竞争的其中一条数据加锁，将该条数据的判重key发送给Redis批量判重单元，处理好后释放锁，其他数据轮询尝试加锁，将加锁成功的数据的判重key发送给Redis批量判重单元，处理后释放锁，循环直至存在竞争的所有数据的判重key均发送给Redis批量判重单元。

Redis批量判重单元，用于Redis批量判断判重key是否存在，若不存在则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1，生成预计算结果。

累加单元，用于将Redis批量判重模块得到预计算结果中相同维度组合下相同度量字段值的度量值进行累加，生成所述实时数据的去重计数结果。

可选的，所述装置还包括：

存储模块，用于将所述实时数据的去重计数结果存储到数据库中。

可选的，所述存储模块进一步用于将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，依次查找数据库中是否存在相同的查找字段，若存在则将相同的查找字段对应的值相加存入数据库，若不存在则将查找字段和其对应的值存入数据库。

本发明实施例提供的实时数据去重计数装置，可实现实时数据的度量去重计数值的精准计算，同时去重过程占用内存空间小。本发明实施例中，将实时数据的所有维度组合分解成1维维度组合，根据1维维度组合生成判重key，使用分布式锁的控制对判重key进行Redis批量去重处理，可以保证各维度组合下实

时数据的度量字段的唯一性，判重得到实时数据的精准去重计数值。同时只需保存1维维度组合的判重key进行判重，相比于现有去重方法，判重key的数量少，大量内存资源得到释放，且判重key的长度固定，容易预估Redis内存使用量。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于设备实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）或随机存储记忆体（Random Access Memory, RAM）等。以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求的保护范围为准。

对说明书中的一些技术名词做出解释如下：

[1] Cube：不同维度的数据组成的数据集。

[2] Cuboid：每一种维度的组合下聚合的数据，Cube的基本单元。

[3] Redis：一个开源的使用ANSI C语言编写、支持网络、可基于内存亦可持久化的日志型、Key-Value数据库。

[4] Kafka：一个高吞吐量的分布式的发布订阅消息系统。

[5] MySQL：一个关系型数据库管理系统。

[6] HBase：建立在Hadoop文件系统之上的分布式的、面向列的开源数据库。

[7] Spark Streaming: 可扩展、容错的流式处理应用程序。

权 利 要 求 书

1、一种实时数据去重计数方法，其特征在于，包括：

获取实时数据和计数信息，所述计数信息包括维度信息、维度组合信息和度量信息；

根据所述维度信息、维度组合信息和度量信息，得到所述实时数据的1维维
5 度组合，生成判重key；

采用分布锁机制对所述判重key进行Redis批量判重，得到所述实时数据的去重计数结果。

2、根据权利要求1所述的方法，其特征在于，采用分布锁机制对所述判重key进行Redis批量判重，得到所述实时数据的去重计数结果，具体包括：

10 判断所述实时数据中各条数据是否存在竞争，若不存在，则所有判重key进行Redis批量判重；若存在，则存在竞争的其中一条数据加锁，数据的判重key进行Redis批量判重，判重结束后释放锁，其他数据轮询加锁，加锁成功的数据的判重key进行Redis批量判重，判重结束后释放锁，循环直至存在竞争的所有数据的判重key均完成批量判重；其中，所述Redis批量判重具体是，Redis批量
15 判断判重key是否存在，若不存在，则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1，生成预计算结果；

将预计算结果中相同维度组合下相同度量字段值的度量值进行累加，生成所述实时数据的去重计数结果。

3、根据权利要求1或2所述的方法，其特征在于，所述方法还包括：将所述
20 实时数据的去重计数结果存储到数据库中。

4、根据权利要求3所述的方法，其特征在于，所述将所述实时数据的去重计数结果存储到数据库中，具体包括：

将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，

依次查找数据库中是否存在相同的查找字段，若存在，则将相同的查找字段对应的值相加后存入数据库；若不存在，则将查找字段和其对应的值存入数据库。

5、根据权利要求1所述的方法，其特征在于，所述维度信息包括维度编码，维度组合信息包括维度组合编码，度量信息包括度量编码。

5 6、根据权利要求5所述的方法，其特征在于，所述生成判重key，具体包括：将1维维度组合的维度编码拼接日期、实时数据的度量字段值和实时数据的维度字段值作为判重key。

7、一种实时数据去重计数装置，其特征在于，包括：

获取模块，用于获取实时数据和计数信息，所述计数信息包括维度信息、

10 维度组合信息和度量信息；

判重key生成模块，用于根据所述维度信息、维度组合信息和度量信息，得到所述实时数据的1维维度组合，生成判重key；

计数模块，用于采用分布锁机制对判重key进行Redis批量判重，得到所述实时数据的去重计数结果。

15 8、根据权利要求7所述的装置，其特征在于，所述计数模块包括：

分布计算单元，用于判断所述实时数据中各条数据是否存在竞争，若不存在，则将所有判重key发送给Redis批量判重单元；若存在，则存在竞争的其中一条数据加锁，将数据的判重key发送给Redis批量判重单元，处理好后释放锁，其他数据轮询加锁，将加锁成功的数据的判重key发送给Redis批量判重单元，

20 处理后释放锁，循环直至存在竞争的所有数据的判重key均发送给Redis批量判重单元；

Redis批量判重单元，用于Redis批量判断判重key是否存在，若不存在则关联该判重key的所有维度组合中不重复的维度组合下的度量值加1，生成预计算

结果；

累加单元，用于将Redis批量判重单元生成的预计算结果中相同维度组合下相同度量字段值的度量值进行累加，生成所述实时数据的去重计数结果。

9、根据权利要求7或8所述的装置，其特征在于，所述装置还包括：

5 存储模块，用于将所述实时数据的去重计数结果存储到数据库中。

10、根据权利要求9所述的装置，其特征在于，所述存储模块进一步用于将所述去重计数结果转换成查找字段和值，所述查找字段与值一一对应，依次查找数据库中是否存在相同的查找字段，若存在则将相同的查找字段对应的值相加后存入数据库，若不存在则将查找字段和其对应的值存入数据库。

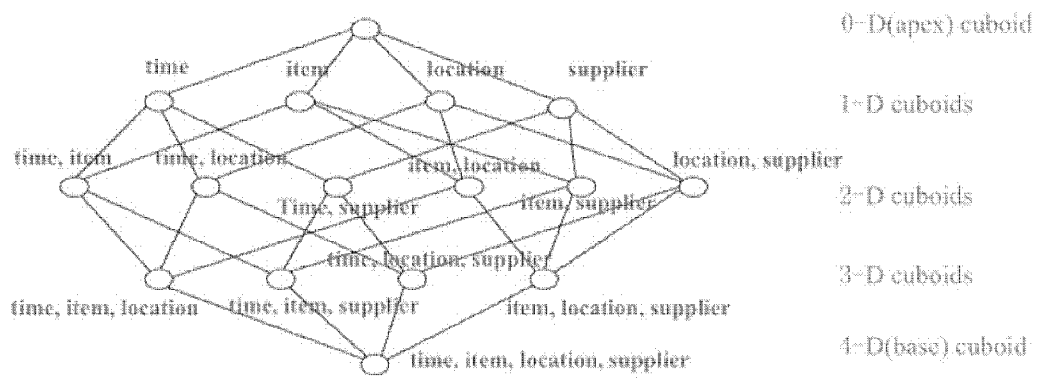


图1

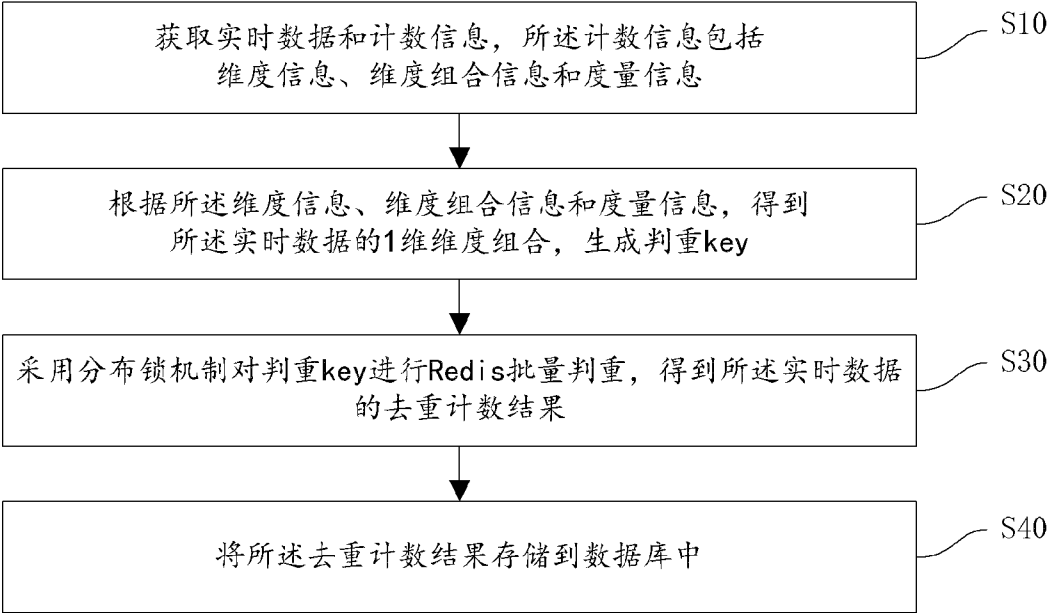


图2

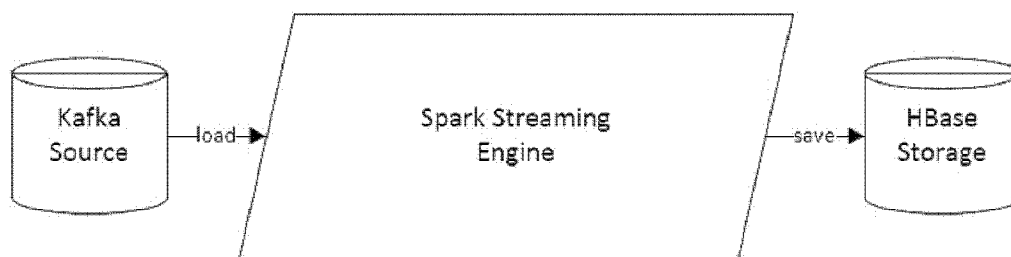


图3

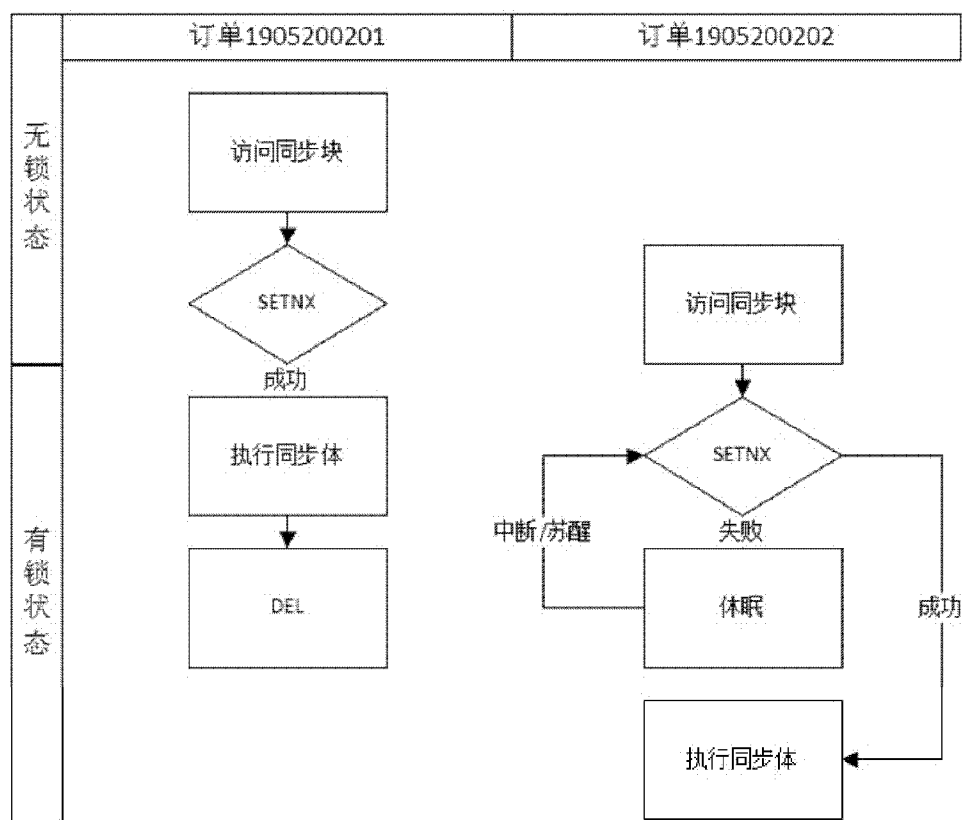


图4

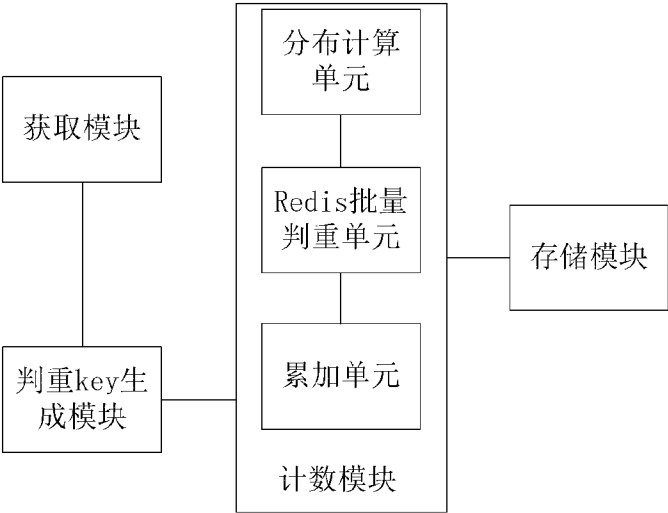


图5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/097839

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, CNPAT, WPI, EPODOC, IEEE: 大数据, 去重, 统计, 计数, 分析, 判重, 关键词, 关键字, 维度, 组合, 分布锁, big data, duplicate removal, statistics, counting, , analysis, duplicate judgment, keywords, dimensionality, combination, cuboid, redis, distribution, lock

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110569263 A (SUNING CLOUD COMPUTING CO., LTD.) 13 December 2019 (2019-12-13) claims 1-10	1-10
A	CN 107665241 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.) 06 February 2018 (2018-02-06) description, paragraphs [0032]-[0074]	1-10
A	CN 105893421 A (LETV INFORMATION TECH (BEIJING) CO., LTD.) 24 August 2016 (2016-08-24) entire document	1-10
A	CN 109816536 A (CHINA PING AN PROPERTY INSURANCE STOCK CO., LTD.) 28 May 2019 (2019-05-28) entire document	1-10
A	CN 109800225 A (BEIJING QIYI CENTURY SCIENCE & TECHNOLOGY CO., LTD.) 24 May 2019 (2019-05-24) entire document	1-10



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

11 September 2020

Date of mailing of the international search report

28 September 2020

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

International application No.

PCT/CN2020/097839

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015213108 A1 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) 30 July 2015 (2015-07-30) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/097839

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110569263	A	13 December 2019	None			
CN	107665241	A	06 February 2018	None			
CN	105893421	A	24 August 2016	None			
CN	109816536	A	28 May 2019	None			
CN	109800225	A	24 May 2019	None			
US	2015213108	A1	30 July 2015	CN	104376020	A	25 February 2015
				WO	2015021850	A1	19 February 2015
				TW	201508524	A	01 March 2015

国际检索报告

国际申请号

PCT/CN2020/097839

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, CNPAT, WPI, EPDOC, IEEE: 大数据, 去重, 统计, 计数, 分析, 判重, 关键词, 关键字, 维度, 组合, 分布锁, big data, duplicate removal, statistics, counting, analysis, duplicate judgment, keywords, dimensionality, combination, cuboid, redis, distribution, lock

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
PX	CN 110569263 A (苏宁云计算有限公司) 2019年 12月 13日 (2019 - 12 - 13) 权利要求1-10	1-10
A	CN 107665241 A (北京京东尚科信息技术有限公司 等) 2018年 2月 6日 (2018 - 02 - 06) 说明书第[0032]-[0074]段	1-10
A	CN 105893421 A (乐视网信息技术北京股份有限公司) 2016年 8月 24日 (2016 - 08 - 24) 全文	1-10
A	CN 109816536 A (中国平安财产保险股份有限公司) 2019年 5月 28日 (2019 - 05 - 28) 全文	1-10
A	CN 109800225 A (北京奇艺世纪科技有限公司) 2019年 5月 24日 (2019 - 05 - 24) 全文	1-10
A	US 2015213108 A1 (TENCENT TECHNOLOGY SHENZHEN COMPANY LIMITED) 2015年 7月 30日 (2015 - 07 - 30) 全文	1-10

☐ 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

2020年 9月 11日

国际检索报告邮寄日期

2020年 9月 28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

授权官员

刘长勇

电话号码 86-(10)-53961368

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/097839

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	110569263	A	2019年 12月 13日	无	
CN	107665241	A	2018年 2月 6日	无	
CN	105893421	A	2016年 8月 24日	无	
CN	109816536	A	2019年 5月 28日	无	
CN	109800225	A	2019年 5月 24日	无	
US	2015213108	A1	2015年 7月 30日	CN 104376020 A	2015年 2月 25日
				WO 2015021850 A1	2015年 2月 19日
				TW 201508524 A	2015年 3月 1日