

(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
2016 年 11 月 10 日 (10.11.2016)



WIPO | PCT



(10) 国际公布号
WO 2016/177279 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/079812
- (22) 国际申请日: 2016 年 4 月 21 日 (21.04.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510222356.4 2015 年 5 月 4 日 (04.05.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼乔治城资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 韩旻 (HAN, Min) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼 阿里巴巴集团法务部, Zhejiang 311121 (CN)。

- (74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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,

[见续页]

(54) Title: DATA PROCESSING METHOD AND SYSTEM

(54) 发明名称: 数据处理的方法及系统

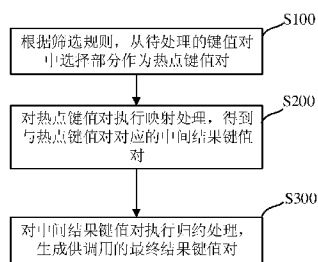


图 1

- S100 Selecting some of the key-value pairs to be processed as hot key-value pairs according to the filtering rules
- S200 Mapping the hot key-value pairs, and obtaining the middle result key-value pairs corresponding to the hot key-value pairs; performing reduction processing on the middle result key-value pairs corresponding to the hot key-value pairs
- S300 Performing reduction processing on the middle result key-value pairs, and generating the final key-value pair to be called

(57) Abstract: Disclosed are a data processing method and system. The method includes the following steps: selecting some of the key-value pairs to be processed as hot key-value pairs according to filtering rules; mapping the hot key-value pairs, and obtaining the middle result key-value pairs corresponding to the hot key-value pairs; performing reduction processing on the middle result key-value pairs, and generating the final key-value pairs to be called, the key-value pairs including the key value characterizing the attribute and the key value characterizing the attribute content. In the method and the corresponding system thereof, the data processing system preprocesses the hot key-values pairs in order to facilitate the business system call, while the non hot key-value pairs are processed in the business system call. Thus the real-time processing data amount of the data processing system, which provides the background service for the business system, is reduced, the efficiency of the data processing is improved, and the waiting time of the business system for data processing results is reduced. The business processing flow is smooth, and then the user experience is good.

(57) 摘要: 本申请公开了一种数据处理的方法及其系统。该方法包括以下步骤: 根据筛选规则, 从待处理的键值对中选择部分作为热点键值对; 对热点键值对执行映射处理, 得到与热点键值对对应的中间结果键值对; 对中间结果键值对执行归约处理, 生成供调用的最终结果键值对; 其中, 键值对包括表征属性的键值和表征属性内容的键值。在该方法及其对应的系统中, 数据处理系统将热点键值对预先处理以便业务系统调用, 而非热点键值对在业务系统调用时才进行处理, 从而, 减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量, 提高数据处理执行效率, 减少了业务系统获得数据处理结果的等待时间, 业务处理顺畅, 进而用户体验好。



WO 2016/177279 A1



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))。

数据处理的方法及系统

技术领域

本申请涉及大数据技术领域，尤其涉及一种数据处理的方法及系统。

5

背景技术

随着计算机技术的发展，计算机需要处理的数据量也越来越大，单台计算机已经不能处理一些大规模的数据，因此发展出将多台计算机联合起来构成计算机群以并行处理大规模数据的技术。

10 Hadoop 分布式集群系统架构即是这样的系统架构。Hadoop 系统可以使用大量廉价的计算机构成计算机集群，通过该集群取代昂贵的高运算速度的计算机进行高速运算和存储。Hadoop 系统主要包括分布式文件系统和映射（Map）归约（Reduce）系统。其中，分布式文件系统进行数据管理和数据存储。映射归约系统实现对分布式文件系统输入的数据进行运算处理，主要包括：将待处
15 理数据集分解为多个数据块；分别对每个数据块中的每个原始键值对数据执行 Map 处理，得到与每个原始键值对数据对应的中间结果键值对数据；在得到与所有原始键值对数据对应的中间结果键值对数据后，再对所有中间结果键值对数据进行相应的归约（Reduce）处理，得到相应的最终结果键值对数据。

通过上述处理方式，可以将一个大的任务拆成较大数量的小任务并由处于
20 分布式系统中的较大数量的计算机（也称为任务执行者）来执行。这样可以完成对海量数据的快速处理。这种处理方式仍然没有减少总的计算资源，而是将需要的大量计算资源分配给了大量的计算机，从而可以大大压缩需要的处理时间。对于对时间敏感度不高的离线场景来说，这种处理方式较为适合。对于在线业务场景，例如即时通讯场景，一般要求在短时间内完成海量数据处理并输
25 出结果，因此对时间较为敏感。

在实现本申请过程中，发明人发现现有技术中至少存在如下问题：

对时间较为敏感的在线业务场景中，仍然是占用大量计算机资源对海量数

据完成处理，即处理的数据量仍然是巨大的。因此，Hadoop 系统处理数据过程消耗的时间冗长，造成业务系统调用 Hadoop 系统对数据处理结果的等待时间长，执行效率低，不能满足具体的业务顺畅的诉求，进而导致用户体验差。

因此，发明人在对现有的数据处理的方法进行研究的基础上，提供一种执行效率高、用户体验好的数据处理的方法及系统。

发明内容

本申请实施例提供一种执行效率高、用户体验好的数据处理的方法，具体的，一种数据处理的方法，包括以下步骤：

- 10 根据筛选规则，从待处理的键值对中选择部分作为热点键值对；
对热点键值对执行映射处理，得到与热点键值对对应的中间结果键值对；
对中间结果键值对执行归约处理，生成供调用的最终结果键值对；
其中，键值对包括表征属性的键值和表征数值的键值。
- 本申请实施例还提供一种数据处理的方法，包括以下步骤：
- 15 对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果键值对；
根据筛选规则，从中间结果键值对中选择部分作为热点键值对；
对热点键值对执行归约处理，生成供调用的最终结果键值对；
其中，键值对包括表征属性的键值和表征属性内容的键值。
- 20 本申请实施例还提供一种数据处理的系统，包括：
筛选模块，用于根据筛选规则，从待处理的键值对中选择部分作为热点键值对；
映射模块，用于对热点键值对执行映射处理，得到与热点键值对对应的中间结果键值对；
25 归约模块，用于对中间结果键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征数值的键值。

本申请实施例还提供一种数据处理的系统，包括：

映射模块，用于对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果键值对；

5 筛选模块，用于根据筛选规则，从中间结果键值对中选择部分作为热点键值对；

归约模块，用于对热点键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

10 本申请实施例提供的数据处理的方法和系统，至少具有如下有益效果：

数据处理系统将热点键值对预先处理以便业务系统调用，而非热点键值对在业务系统调用时才进行处理，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

15

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

20 图 1 为本申请实施例提供的数据处理的方法的流程图。

图 2 为本申请实施例提供的从映射键值对中选择部分作为热点键值对的流程图。

图 3 为本申请实施例提供的数据处理的系统的结构示意图。

25 具体实施方式

为了解决现有的数据处理的方法的数据处理时间长、执行效率低，不能满

足具体的业务顺畅的诉求，进而导致用户体验差的技术问题，本申请实施例提供了一种数据处理的方法及其对应系统。在该方法及其对应的系统中，数据处理系统

将热点键值对预先处理以便业务系统调用，而非热点键值对在业务系统调用时才进行处理，从而，减少了为业务系统提供后台服务的数据处理系统需要

5 要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例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

10 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，都属于本申请保护的范围。

Hadoop 系统可以包括：

客户端 JobClient，用于提交 Map-Reduce 工作；

工作调度器 JobTracker，其为一个 Java 进程，用于协调整个工作运行；

15 任务调度器 TaskTracker，其为一个 Java 进程，运行此工作的任务；

分布式文件系统（Hadoop Distributed File System，HDFS），用于在各个进程间共享工作相关的文件。

Hadoop 系统的工作流程可以包括：

一、 任务提交。

20 客户端向工作调度器请求一个新的工作代码，创建一个新的工作实例，调用 submitJob 函数。

二、 任务初始化。

当工作调度器收到 submitJob 函数调用的时候，获取并初始化任务。工作调度器创建任务，并分配任务代码。

25 三、 任务分配。

工作调度器给任务调度器分配任务。

四、 任务执行。

任务调度器被分配了一个任务后，开始运行此任务。映射过程中，任务调度器调用 map 函数处理任务，也即处理原始键值对生成中间结果键值对，并将中间结果键值对按照键值的顺序输出。然后，任务调度器调用 reduce 函数处
5 理中间结果键值对生成最终结果键值对。

五、 任务结束

工作调度器获得任务调度器的所有任务运行成功的报告后，结束工作。

图 1 为本申请实施例提供的数据处理的方法的流程图，具体包括以下步骤：
骤：

10 S100：根据筛选规则，从待处理的键值对中选择部分作为热点键值对。

数据体现为描述数据性质的属性及数值，也就是通常讲的键值对。键值对包括表征属性的键值和表征属性内容的键值。属性内容包括但不限于列表 list、哈希表 hash map、字符串、数值、布尔值、有序列表 array、null 值等。例如，
{“姓名”：“王小二”}表示人物的“姓名”为“王小二”的数据。

15 在一个具体实施例中，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：可以随机选择若干待处理的键值对作为热点键值对。实际上，判断待处理的键值对是否是热点键值对是一个复杂的过程，尤其是待处理键值对的数量级在百万，甚至上亿的情形下。在本申请实施例中，数据处理系统随机选择若干待处理键值对作为热点键值对，从而可以简化待处理键值
20 对是否是热点键值对的判断过程，提高使用该方法的数据处理效率。

将热点键值对预先处理供业务系统调用，而非热点键值对在业务系统调用时才进行处理，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

25 在又一个具体实施例中，请参照图 2，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：

S101: 随机选择第一数量的待处理的键值对作为待选键值对;

S102: 统计待选键值对中每一键值对被调用的频度;

S103: 按照频度排列待选键值对;

S104: 从待选键值对中选择第二数量的、调用的频度位于前列的键值对作为
5 热点键值对;

其中, 第一数量大于第二数量。

在本申请实施例中, 首先, 数据处理系统随机选择第一数量的映射键值对作为待选键值对。第一数量通常与具体的业务相对应。该第一数量的数值可以是根据历史经验设定的固定值, 也可以是计算机动态调整变化生成的数值。

10 然后, 数据处理系统统计待选键值对中每一键值对被调用的频度, 按照频度排列待选键值对。在具体的业务活动中, 往往需要调取一定数量的键值对来对业务系统进行支持。此时, 数据处理系统对每一键值对被调用的频度——一段时间内被调用的次数——进行追踪记录。进一步的, 数据处理系统还可以按照被调用的频度从大到小排列键值对。

15 接着, 数据处理系统从待选键值对中选择第二数量的、调用的频度位于前列的键值对作为热点键值对。其中, 第一数量大于第二数量。同样, 该第二数量的数值可以是根据历史经验设定的固定值, 也可以是数据处理系统动态调整变化生成的数值。选择出来的热点键值对相对于其他的待选键值对被调用的频度大。数据处理系统预先将热点键值对处理, 而不是对其他的待选键值对进行
20 预先处理。预先处理过的键值对被调用的概率大于其他键值对被调用的概率。从而, 减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量, 提高数据处理执行效率, 减少了业务系统获得数据处理结果的等待时间, 业务处理顺畅, 进而用户体验好。

进一步的, 在本申请实施例中, 从映射键值对中选择部分作为热点键值对,
25 还包括:

在随机选择第一数量的待处理的键值对作为待选键值对的步骤之前, 设定

待选键值对的业务类别条件集；

选择符合业务类别条件集的待处理的键值对。

在本申请实施例中，业务类别条件集可以根据历史经验设定的固定，也可以动态调整变化生成。实际上，一种业务活动的业务系统调用的键值对往往带有一些特有的性质以区别于其他业务活动。例如，推送信息的业务系统调用的键值对，相较于支付的业务系统调用的键值对，有其特有的性质。推送信息的业务系统，可以与表示接收者年龄的键值对相关。例如，婚庆商品的推送信息对于年龄在 16 岁以下的接收者通常属于垃圾信息。推送信息的业务系统的业务类别条件集包含表示年龄的键值对时，可以取得较好的推送效果。

因此，设定待处理的键值对的业务类别条件集，数据处理系统经过业务类别条件集的判断可以过滤掉大量的待处理的键值对，提高热点键值对的选择精确度，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

在本申请的又一个具体实施例中，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：

设定热点键值对的调用频度阈值；

当键值对被调用的频度大于调用频度阈值时，将该键值对设置为热点键值对。

在本申请实施例中，设定键值对的调用频度阈值，当键值对的被调用的频度大于调用频度阈值时，数据处理系统将该键值对设置为热点键值对。数据处理系统预先将热点键值对处理，而不是对其他的键值对进行预先处理。预先处理过的键值对被调用的概率大于其他键值对被调用的概率。从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

S200: 对热点键值对执行映射处理, 得到与热点键值对对应的中间结果键值对。

在本申请提供的一个实施例中, 客户端 JobClient 向工作调度器提交 Map-Reduce 工作, 创建一个新的工作实例, 调用 submitJob 函数。当工作调度器收到 submitJob 函数调用的时候, 获取并初始化任务。工作调度器创建一个任务, 并分配一个任务代码。工作调度器给任务调度器分配任务。任务调度器被分配了一个任务后, 开始运行此任务。映射过程中, 任务调度器调用 map 函数处理任务, 也即处理原始键值对生成中间结果键值对, 并将中间结果键值对按照键值的顺序输出。

S300: 对中间结果键值对执行归约处理, 生成供调用的最终结果键值对。

在本步骤中, 任务调度器调用 reduce 函数处理中间结果键值对生成最终结果键值对。工作调度器获得任务调度器的所有任务运行成功的报告后, 将最终结果键值对存入分布式文件系统, 并结束工作。

在本申请实施例中, 数据处理系统将热点键值对预先处理以便业务系统调用, 而非热点键值对在业务系统调用时才进行处理, 从而, 减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量, 提高数据处理执行效率, 减少了业务系统获得数据处理结果的等待时间, 业务处理顺畅, 进而用户体验好。

在本申请提供的一个实施例中, 方法还包括:

使用机器学习模型, 优化对筛选规则。

机器学习模型, 涉及人工智能。在本申请实施例中, 使用机器学习模型对筛选规则进行优化。数据处理系统运行一段时间后, 热点键值对和非热点键值对判断的准确性可以显著提高。从而, 减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量, 提高数据处理执行效率, 减少了业务系统获得数据处理结果的等待时间, 业务处理顺畅, 进而用户体验好。

使用何种机器学习模型, 根据具体的业务系统来进行选择, 下面简单介绍

一种机器学习模型对筛选规则的优化。

具体的，使用机器学习模型中的聚类算法，统计键值对被调用的频度随单一属性的分布状况；

根据键值对被调用的频度随单一属性的分布状况，选择键值对被调用的频度不小于预设频度阈值的属性内容的键值的区间；

设定所述属性内容的键值的区间为筛选规则的规则条件。

仍以上面的推送信息的业务系统为例来说明。假设，业务系统统计出推送信息的业务超出某一预设比例，例如 50%，调用表示接收者年龄的键值对。机器学习模型通过 K-means 聚类算法对筛选规则进行优化。

假设要把样本集（表示接收者年龄的键值对及其被调用的频度）分为 m 个类别（频度段），算法描述如下：

（1）适当选择 m 个类别（频度段）的初始中心（频度）；

（2）在第 k 次迭代中，对任意一个样本（表示接收者年龄的键值对及其被调用的频度），求其到 m 个中心的距离（频度差），将该样本（表示接收者年龄的键值对及其被调用的频度）归到距离最短的中心所在的类别（频度段）；

（3）利用均值方法更新该类别（频度段）的中心值（频度）；

（4）对于所有的 m 个中心值（频度），如果利用（2）（3）的迭代法更新后，值保持不变，则迭代结束，否则继续迭代；

（5）针对 m 个类别（频度段）中的每一个类别（频度段）适当选择 n 个类别（年龄段）的初始中心（年龄）；

（6）在第 k 次迭代中，对任意一个样本（表示接收者年龄的键值对及其被调用的频度），求其到 n 个中心的距离（年龄差），将该样本（表示接收者年龄的键值对及其被调用的频度）归到距离最短的中心所在的类别（年龄段）；

（7）利用均值方法更新该类别（年龄段）的中心值（年龄）；

（8）对于所有的 n 个中心值（年龄），如果利用（6）（7）的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

通过该算法，可以计算出调用的频度大的待处理的键值对随年龄的聚类规律。将接收者年龄位于某个类别（年龄段）作为筛选规则的规则条件，例如，将表示接收者年龄位于 12-18 岁作为待处理的键值对是热点键值对的规则条件。业务系统通过机器学习模型对筛选规则优化后，则根据优化后的筛选规则，
5 将热点键值对从待处理的键值对中筛选出。

在本申请提供的实施例中，规则优化模块，还用于：

当一个属性的键值对和另一属性的键值对被具有同一业务代码的业务系统调用时，将这两个属性的键值对的属性内容的键值的区间的并集设定为筛选规则的规则条件。

10 假设，机器学习模型完成接收者年龄这一维度对筛选规则的优化后，还完成了接收者职业这一维度对筛选规则的优化。

数据处理系统计算出接收者处于某一年龄段的待处理的键值对和接收者属于某一职业的待处理的键值对，与推送信息高度相关。例如，某待处理的键值对表示接收者处于 20-30 年龄段，某待处理的键值对表示接收者处于计算机
15 行业，业务系统推送信息时，对同时具有这两个维度特征的接收者能获得良好的业务推广效果。则，机器学习模型将表示接收者处于 20-30 年龄段的待处理的键值对，和表示接收者处于计算机行业的待处理的键值对进行关联形成热点键值对数据组。

数据处理系统通过机器学习模型进一步对热点键值对数据组被调用的频
20 度进行排序，将热点键值对数据组区分为热点数据组和非热点数据组。热点数据组的一种动态调整模式为：设定热点数据组的调用的频度阈值，当数据组中键值对被调用的频度大于频度阈值时，将该数据组设定为热点数据组。

在本申请实施例中，设定数据组处理优先级值。该优先级值由待处理的键值对的加权和值算得。根据该优先级值，动态调整数据组的处理优先级。当数
25 据组中某一键值对被调用一次，数据组的优先级值增加一个单位。当某一数据组的优先级值超过其前一个数据组的优先级值时，数据处理系统将该数据组前

置一位。经过机器学习模型对筛选规则的优化，数据处理系统从待处理的键值对中选择出的热点键值对是被调用频度位于前列的映射键值对，其中由键值对关联形成的热点数据组是被调用频度位于前列的数据组，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

在本申请提供的一个实施例中，方法还包括：

当非热点键值对被调用时，使用归约函数处理非热点键值对生成供调用的数据。

在本申请实施例中，热点键值对由数据处理系统使用归约函数预先处理生成供业务系统调用的数据；非热点键值对在被业务系统调用时，由数据处理系统使用归约函数即时处理，生成供业务系统调用的数据。从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

在本申请实施例的一种替代方式中，数据处理的方法，包括以下步骤：

对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果键值对；

根据筛选规则，从中间结果键值对中选择部分作为热点键值对；

对热点键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

需要指出的是，与前面提供的具体实施例的差别在于：选择部分作为热点键值对的步骤，设置于映射处理步骤之后。在本申请实施例中，降低了归约处理的键值对的数据量，可以在一定程度上解决数据量巨大的问题，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，

进而用户体验好。

以上是本申请实施例提供的数据处理的方法，基于同样的思路，请参照图3，本申请还提供一种数据处理的系统1，包括：

5 筛选模块10，用于根据筛选规则，从待处理的键值对中选择部分作为热点键值对；

映射模块20，用于对热点键值对执行映射处理，得到与热点键值对对应的中间结果键值对；

归约模块30，用于对中间结果键值对执行归约处理，生成供调用的最终结果键值对；

10 其中，键值对包括表征属性的键值和表征数值的键值。

进一步的，筛选模块10，用于从待处理的键值对中选择部分作为热点键值对，具体用于：

随机选择若干待处理的键值对作为热点键值对。

15 进一步的，筛选模块10，用于从待处理的键值对中选择部分作为热点键值对，具体用于：

随机选择第一数量的待处理的键值对作为待选键值对；

统计待选键值对中每一键值对被调用的频度；

按照频度排列待选键值对；

20 从待选键值对中选择第二数量的、调用的频度位于前列的键值对作为热点键值对；

其中，第一数量大于第二数量。

进一步的，筛选模块10，用于从映射键值对中选择部分作为热点键值对，具体还用于：

25 在随机选择第一数量的待处理的键值对作为待选键值对的步骤之前，设定待选键值对的业务类别条件集；

选择符合业务类别条件集的待处理的键值对。

进一步的，筛选模块 10，用于从待处理的键值对中选择部分作为热点键值对，具体用于：

设定热点键值对的调用频度阈值；

当键值对被调用的频度大于调用频度阈值时，将该键值对设置为热点键

5 对。

进一步的，系统还包括规则优化模块 40，用于：

使用机器学习模型，优化对筛选规则。

进一步的，映射模块 20，用于：

当非热点键值对被调用时，对非热点键值对执行映射处理，得到与非热点
10 键值对对应的中间结果键值对。

进一步的，一种数据处理系统 1，包括：

映射模块 20，用于对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果键值对；

筛选模块 10，用于根据筛选规则，从中间结果键值对中选择部分作为热
15 点键值对；

归约模块 30，用于对热点键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

在本申请实施例中，数据处理系统将热点键值对预先处理以便业务系统调
20 用，而非热点键值对在业务系统调用时才进行处理，从而，减少了为业务系统提供后台服务的数据处理系统需要即时处理的数据量，提高数据处理执行效率，减少了业务系统获得数据处理结果的等待时间，业务处理顺畅，进而用户体验好。

本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计
25 算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包

含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数值处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数值处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数值处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数值处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数值结构、程序的模块或其他数值。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其

他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带, 磁带磁
5 磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算
设备访问的信息。按照本文中的界定, 计算机可读介质不包括暂存电脑可读媒
体 (transitory media), 如调制的数值信号和载波。

还需要说明的是, 术语“包括”、“包含”或者其任何其他变体意在涵盖非排
他性的包含, 从而使得包括一系列要素的过程、方法、商品或者设备不仅包括
那些要素, 而且还包括没有明确列出的其他要素, 或者是还包括为这种过程、
10 方法、商品或者设备所固有的要素。在没有更多限制的情况下, 由语句“包括
一个……”限定的要素, 并不排除在包括所述要素的过程、方法、商品或者设
备中还存在另外的相同要素。

本领域技术人员应明白, 本申请的实施例可提供为方法、系统或计算机程
序产品。因此, 本申请可采用完全硬件实施例、完全软件实施例或结合软件和
15 硬件方面的实施例的形式。而且, 本申请可采用在一个或多个其中包含有计算
机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、
光学存储器等)上实施的计算机程序产品的形式。

以上所述仅为本申请的实施例而已, 并不用于限制本申请。对于本领域技
术人员来说, 本申请可以有各种更改和变化。凡在本申请的精神和原理之内所
20 作的任何修改、等同替换、改进等, 均应包含在本申请的权利要求范围之内。

权 利 要 求 书

1、一种数据处理的方法，其特征在于，包括以下步骤：

根据筛选规则，从待处理的键值对中选择部分作为热点键值对；

对热点键值对执行映射处理，得到与热点键值对对应的中间结果键值对；

5 对中间结果键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

2、如权利要求 1 所述的方法，其特征在于，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：

随机选择若干待处理的键值对作为热点键值对。

10 3、如权利要求 1 所述的方法，其特征在于，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：

随机选择第一数量的待处理的键值对作为待选键值对；

统计待选键值对中每一键值对被调用的频度；

按照频度排列待选键值对；

15 从待选键值对中按照调用频度由高到低的顺序选择第二数量的键值对作为热点键值对；

其中，第一数量大于第二数量。

4、如权利要求 3 所述的方法，其特征在于，所述根据筛选规则，从映射键值对中选择部分作为热点键值对，还包括：

20 在随机选择第一数量的待处理的键值对作为待选键值对的步骤之前，设定待选键值对的业务类别条件集；

选择符合业务类别条件集的待处理的键值对。

5、如权利要求 1 所述的方法，其特征在于，根据筛选规则，从待处理的键值对中选择部分作为热点键值对，具体包括：

25 设定热点键值对的调用频度阈值；

当键值对被调用的频度大于调用频度阈值时，将该键值对设置为热点键值

对。

6、如权利要求 1 所述的方法，其特征在于，所述方法还包括：

使用机器学习模型，优化筛选规则。

7、如权利要求 6 所述的方法，其特征在于，使用机器学习模型，优化筛

5 选规则，具体包括：

使用机器学习模型中的聚类算法，统计键值对被调用的频度随单一属性的分布状况；

根据键值对被调用的频度随单一属性的分布状况，选择键值对被调用的频度不小于预设频度阈值的属性内容的键值的区间；

10 设定所述属性内容的键值的区间为筛选规则的规则条件。

8、如权利要求 7 所述的方法，其特征在于，所述方法还包括：

当一个属性的键值对和另一属性的键值对被具有同一业务代码的业务系统调用时，将这两个属性的键值对的属性内容的键值的区间的并集设定为筛选规则的规则条件。

15 9、如权利要求 1 所述的方法，其特征在于，所述方法还包括：

当非热点键值对被调用时，使用归约函数处理非热点键值对生成供调用的数据。

10、一种数据处理的方法，其特征在于，包括以下步骤：

对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果

20 键值对；

根据筛选规则，从中间结果键值对中选择部分作为热点键值对；

对热点键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

11、一种数据处理的系统，其特征在于，包括：

25 筛选模块，用于根据筛选规则，从待处理的键值对中选择部分作为热点键值对；

映射模块，用于对热点键值对执行映射处理，得到与热点键值对对应的中间结果键值对；

归约模块，用于对中间结果键值对执行归约处理，生成供调用的最终结果键值对；

5 其中，键值对包括表征属性的键值和表征数值的键值。

12、如权利要求 11 所述的系统，其特征在于，所述筛选模块，用于从待处理的键值对中选择部分作为热点键值对，具体用于：

随机选择若干待处理的键值对作为热点键值对。

13、如权利要求 11 所述的系统，其特征在于，所述筛选模块，用于从待
10 处理的键值对中选择部分作为热点键值对，具体用于：

随机选择第一数量的待处理的键值对作为待选键值对；

统计待选键值对中每一键值对被调用的频度；

按照频度排列待选键值对；

从待选键值对中选择第二数量的、调用的频度位于前列的键值对作为热点
15 键值对；

其中，第一数量大于第二数量。

14、如权利要求 13 所述的系统，其特征在于，所述筛选模块，用于从映射键值对中选择部分作为热点键值对，具体还用于：

在随机选择第一数量的待处理的键值对作为待选键值对的步骤之前，设定
20 待选键值对的业务类别条件集；

选择符合业务类别条件集的待处理的键值对。

15、如权利要求 11 所述的系统，其特征在于，所述筛选模块，用于从待处理的键值对中选择部分作为热点键值对，具体用于：

设定热点键值对的调用频度阈值；

25 当键值对被调用的频度大于调用频度阈值时，将该键值对设置为热点键值对。

16、如权利要求 11 所述的系统，其特征在于，所述系统还包括规则优化模块，用于：

使用机器学习模型，优化对筛选规则。

17、如权利要求 16 所述的系统，其特征在于，所述规则优化模块，用于：

5 使用机器学习模型中的聚类算法，统计键值对被调用的频度随单一属性的分布状况；

根据键值对被调用的频度随单一属性的分布状况，选择键值对被调用的频度不小于预设频度阈值的属性内容的键值的区间；

设定所述属性内容的键值的区间为筛选规则的规则条件。

10 18、如权利要求 17 所述的系统，其特征在于，所述规则优化模块，还用于：

当一个属性的键值对和另一属性的键值对被具有同一业务代码的业务系统调用时，将这两个属性的键值对的属性内容的键值的区间的并集设定为筛选规则的规则条件。

15 19、如权利要求 18 所述的系统，其特征在于，当非热点键值对被调用时，所述映射模块，用于：

对非热点键值对执行映射处理，得到与非热点键值对对应的中间结果键值对。

20、一种数据处理系统，其特征在于，包括：

20 映射模块，用于对待处理的键值对执行映射处理，得到与待处理的键值对对应的中间结果键值对；

筛选模块，用于根据筛选规则，从中间结果键值对中选择部分作为热点键值对；

25 归约模块，用于对热点键值对执行归约处理，生成供调用的最终结果键值对；

其中，键值对包括表征属性的键值和表征属性内容的键值。

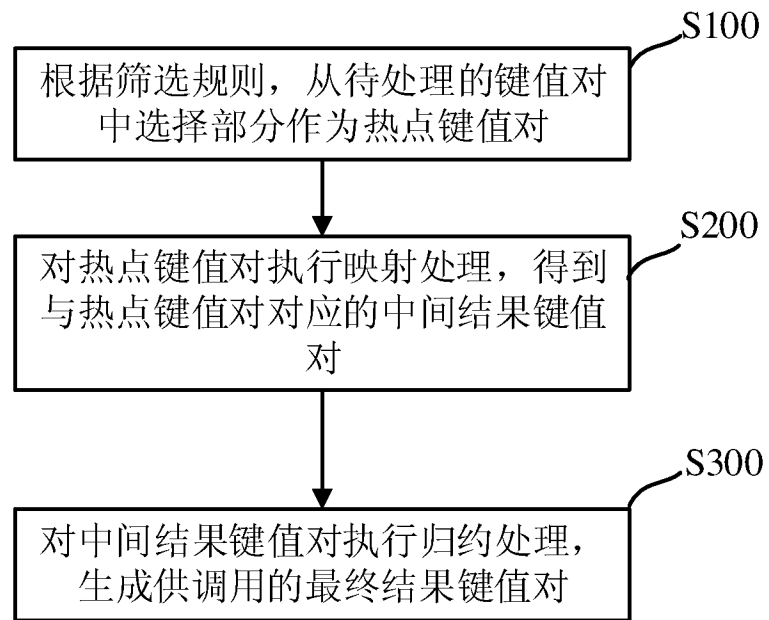


图 1

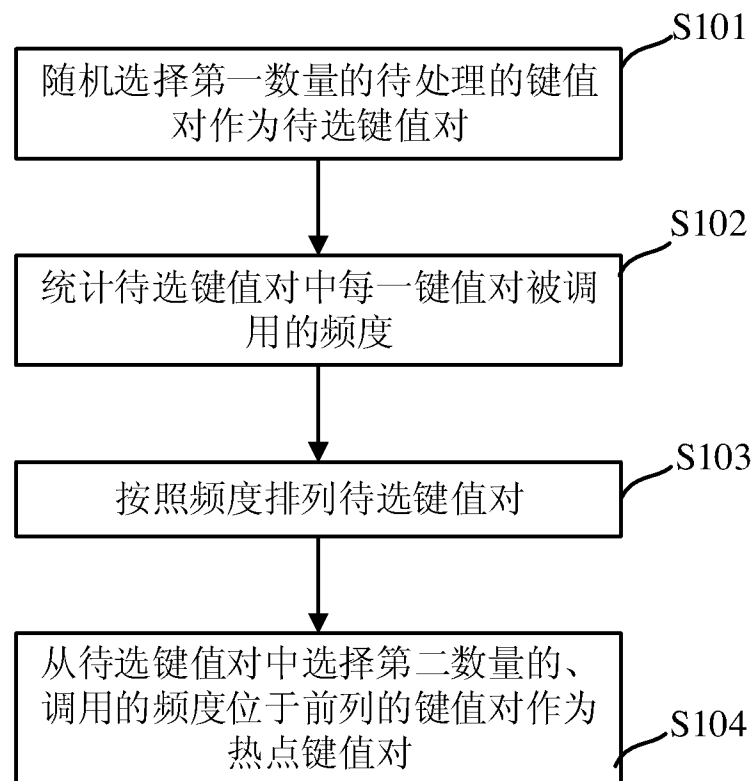


图 2

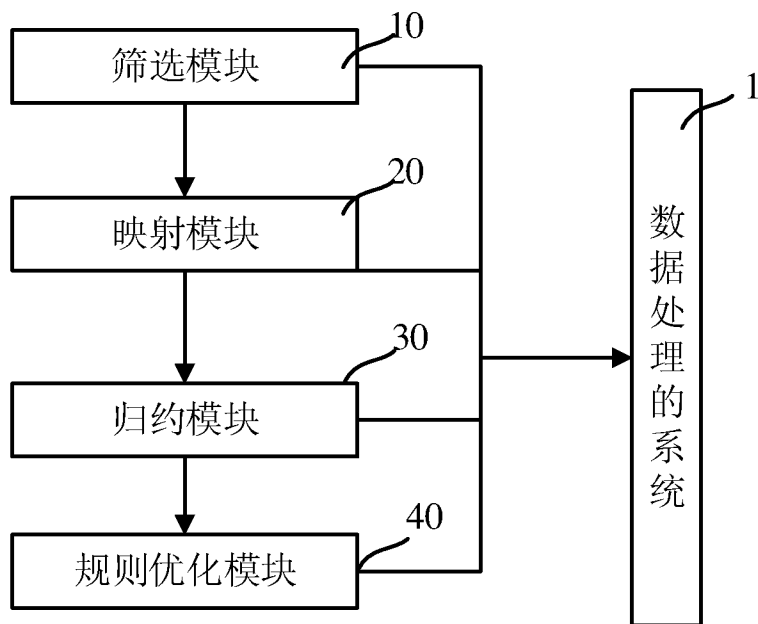


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/079812

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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 17/-

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)

CNXTX; CNABS; VEN; CNKI: key-value pair, protocol, simplification, screening, key value, map????, reduc????, select, choose, filter, distributed, cluster

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102456031 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.), 16 May 2012 (16.05.2012), description, paragraphs [0061]-[0090], and figures 1, 3 and 8	1-20
A	CN 104331464 A (XJ ELECTRIC CO., LTD. et al.), 04 February 2015 (04.02.2015), the whole document	1-20
A	CN 102314336 A (TENCENT INC.), 11 January 2012 (11.01.2012), the whole document	1-20
A	CN 102236581 A (INTERNATIONAL BUSINESS MACHINES CORPORATION), 09 November 2011 (09.11.2011), the whole document	1-20
A	US 7756919 B1 (GOOGLE INC.), 13 July 2010 (13.07.2010), the whole document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 20 June 2016 (20.06.2016)	Date of mailing of the international search report 30 June 2016 (30.06.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer HAO, Xiaoli Telephone No.: (86-10) 62089281

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/079812

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102456031 A	16 May 2012	None	
CN 104331464 A	04 February 2015	None	
CN 102314336 A	11 January 2012	CN 102314336 B	13 April 2016
CN 102236581 A	09 November 2011	US 2013104135 A1	25 April 2013
		EP 2564318 A1	06 March 2013
		WO 2011134875 A1	03 November 2011
		CN 102236581 B	14 August 2013
US 7756919 B1	13 July 2010	US 2014096138 A1	03 April 2014
		US 2010122065 A1	13 May 2010
		US 8612510 B2	17 December 2013

国际检索报告

国际申请号

PCT/CN2016/079812

A. 主题的分类

G06F 17/30(2006.01)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)

G06F17/-

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))

CNTXT;CNABS;VEN;CNKI:键值对, 映射, 归约, 规约, 化简, 选择, 筛选, 分布式, 集群; key value, map????, reduc????, select, choose, filter, distributed, cluster

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 102456031 A (腾讯科技深圳有限公司) 2012年 5月 16日 (2012 - 05 - 16) 说明书第[0061]-[0090]段、附图1, 3, 8	1-20
A	CN 104331464 A (许继电气股份有限公司等) 2015年 2月 4日 (2015 - 02 - 04) 全文	1-20
A	CN 102314336 A (深圳市腾讯计算机系统有限公司) 2012年 1月 11日 (2012 - 01 - 11) 全文	1-20
A	CN 102236581 A (国际商业机器公司) 2011年 11月 9日 (2011 - 11 - 09) 全文	1-20
A	US 7756919 B1 (GOOGLE INC) 2010年 7月 13日 (2010 - 07 - 13) 全文	1-20

☐ 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

2016年 6月 20日

国际检索报告邮寄日期

2016年 6月 30日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(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

郝晓丽

电话号码 (86-10)62089281

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/079812

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	102456031	A	2012年 5月 16日	无			
CN	104331464	A	2015年 2月 4日	无			
CN	102314336	A	2012年 1月 11日	CN	102314336	B	2016年 4月 13日
CN	102236581	A	2011年 11月 9日	US	2013104135	A1	2013年 4月 25日
				EP	2564318	A1	2013年 3月 6日
				WO	2011134875	A1	2011年 11月 3日
				CN	102236581	B	2013年 8月 14日
US	7756919	B1	2010年 7月 13日	US	2014096138	A1	2014年 4月 3日
				US	2010122065	A1	2010年 5月 13日
				US	8612510	B2	2013年 12月 17日

表 PCT/ISA/210 (同族专利附件) (2009年7月)