

(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



(10) 国际公布号
WO 2020/118928 A1

(43) 国际公布日
2020 年 6 月 18 日 (18.06.2020)

(51) 国际专利分类号:
G06F 16/2458 (2019.01)

(21) 国际申请号: PCT/CN2019/077741

(22) 国际申请日: 2019 年 3 月 12 日 (12.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811510849.8 2018 年 12 月 11 日 (11.12.2018) CN

(71) 申请人: 东北大学 (NORTHEASTERN UNIVERSITY) [CN/CN]; 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。

(72) 发明人: 徐泉(XU, Quan); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。 丁

进良(DING, Jinliang); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (CN)。

(74) 代理人: 沈阳东大知识产权代理有限公司 (SHENYANG DONGDA INTELLECTUAL PROPERTY AGENCY CO., LTD); 中国辽宁省沈阳市和平区文化路三巷 11 号, Liaoning 110819 (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, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

(54) Title: DISTRIBUTED TIME SEQUENCE PATTERN RETRIEVAL METHOD FOR MASSIVE EQUIPMENT OPERATION DATA

(54) 发明名称: 一种面向海量设备运行数据的分布式时间序列模式检索方法

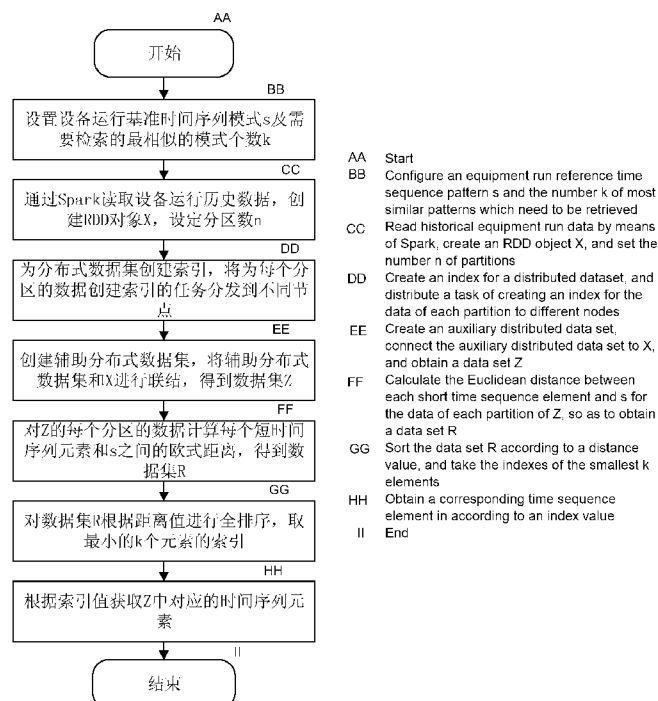


图 1

(57) Abstract: A distributed time sequence pattern retrieval method for massive equipment operation data, which relates to the technical field of time sequence analysis. The method comprises: first, configuring a retrieval reference time sequence pattern and the number of the most similar patterns which need to be retrieved; reading run time sequence data of historical equipment to be retrieved into a distributed system so as to construct a distributed data set X; then, constructing an index for elements of X; constructing multiple auxiliary distributed data sets according to the length of a reference time sequence, connecting X and an auxiliary data set, and constructing a

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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

— 关于申请人有权申请并被授予专利(细则
4.17(ii))

本国际公布:

— 包括国际检索报告(条约第21条(3))。

distributed data set Z, each element thereof being a short time sequence; calculating the Euclidean distance between the reference time sequence and the elements of the Z, so as to construct a distributed data set R; sorting R, taking the smallest k elements, and returning the index of a corresponding element; and obtaining the corresponding element in the data set Z according to the index. The present method improves the similarity retrieval efficiency of a large amount of equipment running time sequence data.

(57) 摘要: 一种海量设备运行数据的分布式时间序列模式检索方法, 涉及时间序列分析技术领域。该方法首先设置检索基准时间序列模式及需要检索的最相似的模式个数; 将待检索的历史设备运行时间序列数据读取进分布式系统中, 构造分布式数据集X; 然后为X的元素构建索引; 根据基准时间序列的长度构造多个辅助分布式数据集, 将X和辅助数据集进行联结, 构造分布式数据集Z, 其每个元素为一个短时间序列; 计算基准时间序列和Z的元素之间的欧式距离, 构造出分布式数据集R; 对R进行排序, 取最小的k个元素, 返回对应元素的索引; 根据索引取得数据集Z中相应的元素。该方法提高了海量设备运行时间序列数据的相似性检索效率。

一种面向海量设备运行数据的分布式时间序列模式检索方法

技术领域

本发明涉及时间序列分析技术领域，尤其涉及一种海量设备运行数据的分布式时间序列模式检索方法。

背景技术

时间序列是一种和时间相关联、具有先后次序的数值序列或符号序列，在金融、气象、故障诊断等领域应用非常广泛。设备运行数据作为设备日常运维数据的重要组成部分，具有典型的时间序列特征，对时间序列的分析处理也是目前设备运行故障诊断研究的一个重要方向，包括异常模式的检索、模式挖掘、聚类等。由于大型设备传感器数量多、采集频率快，因此一般大型设备运行数据具有数据量大、维度高、更新频率高等特点，即具有典型的大数据特征。因此如何对设备海量运行数据组成的庞大的时间序列的高效处理，是目前需要研究和待解决的问题。

时间序列的相似模式检索问题可描述为给定某时间序列模式，从一个大型的时间序列中找出与之最相似的若干子序列，时间序列相似性检索是其他时间序列分析任务实现的前提，比如异常检测和模式识别等。目前的时间序列相似模式检索方法主要采用单机方法对时间序列进行串行的检索，找出相似性符合要求的所有子序列模式。但是由于机器性能的限制，单机方法能够处理的数据量有限，且计算的效率低下，难以满足海量设备运行时间序列数据的检索需求。

目前大数据和云计算的发展，使得数据的分布式并行计算成为了可能，极大的提高了数据处理的能力和效率，为海量设备运行时间序列数据的分析问题提供了解决的思路。为了提高海量设备运行时间序列数据的检索效率，需要研究分布式的时间序列相似模式检索方法。

发明内容

本发明要解决的技术问题是针对上述现有技术的不足，提供一种海量设备运行数据的分布式时间序列模式检索方法，使用分布式计算节点组成的大数据处理集群，通过将数据分发到集群不同节点的内存中，当集群运行计算任务时，将计算任务分解分发到不同的节点，从而实现时间序列相似模式的并行检索。

为解决上述技术问题，本发明所采取的技术方案是：一种海量设备运行数据的分布式时间序列模式检索方法，包括以下步骤：

步骤 1、设置检索的设备运行数据基准时间序列模式 s ，确定检索模式的长度 m ，设定需要检索的最相似的模式个数为 k ；

所述设备运行数据，采样周期为 T 秒；

所述基准时间序列模式是在设备运行数据中自定义选取的一个连续时间段的数据，长度为 m ；

步骤 2、将待检索的海量历史设备运行时间序列数据读取进分布式系统计算节点的内存中，构造初始分布式数据集 X ，并确定并行计算任务的个数 n ，具体包括以下步骤：

步骤 2.1、将保存在硬盘存储介质中的历史设备运行数据通过网络上传至分布式文件系统 HDFS（Hadoop Distributed File System）中；

步骤 2.2、使用分布式 Spark 计算引擎读取存储在 HDFS 的海量设备运行时间序列数据；

步骤 2.3、设置要创建的分布式数据集的分区个数为 n ，Spark 计算引擎根据设定的分区数将存储在 HDFS 上的设备运行数据切分出 n 个数据块，为每个数据块创建一个分区，构造出分区数为 n 的分布式数据集 RDD（Resilient Distributed DataSet，即弹性分布式数据集）对象 X ， X 的每个元素为某一时刻的设备运行数据值，并且 X 维护着分区的顺序及每个分区首个元素在整个数据集中的偏移量；

步骤 2.4、根据分区数确定并行计算任务个数为 n ，Spark 计算引擎为 RDD 对象 X 的每个数据分区创建一个计算任务分发到不同计算节点，任务之间相互独立，每个任务并行处理一个分区的数据；

步骤 3、为步骤 2 中分布式数据集 X 构建索引，根据时间顺序对数据集 X 中的每个元素从 0 开始编号，其中每个分区的编号任务是在不同节点并行进行计算的；数据集 X 中的每个元素转换为 $\langle key, value \rangle$ 键值对记录形式，其中， key 为索引编号， $value$ 为对应时刻的设备运行时间序列数值；

步骤 4、根据基准时间序列模式的长度构造 $m-1$ 个辅助分布式数据集 Y_j ，其中 $j \in (1, m-1)$ ，将步骤 3 中分布式数据集 X 和辅助分布式数据集进行联结，构造出每个元素是一个长度为 m 的短时间序列的分布式数据集 Z ，具体步骤包括：

步骤 4.1、根据设定的基准时间序列模式的长度 m ，构造出 $m-1$ 个和步骤 3 中数据集 X 相同的辅助分布式数据集 Y_j ，其中 $j \in (1, m-1)$ ；

步骤 4.2、对 $m-1$ 个辅助分布式数据集 Y_j 构建索引，其中 Y_j 的元素从 $-j$ 开始编号；

步骤 4.3、对分布式数据集 X 依次和 $m-1$ 个辅助分布式数据集 Y_j 进行笛卡尔积操作，并将进行笛卡尔积的两个分布式数据集中具有相同索引 key 值的元素进行联结；

步骤 4.4、经过步骤 4.3 之后，从步骤 3 中的数据集合 X 构造出分布式数据集 Z ， Z 的元素为 $\langle key_1, (value_1, value_2, \dots, value_m) \rangle$ 形式，其中， key_1 为元素的编号， $(value_1, value_2, \dots, value_m)$ 代表一段以第 key_1 个时刻开始长度为 m 的短时间序列；

步骤 5、通过 Spark 计算引擎创建计算任务，每个计算任务的逻辑是计算基准时间序列模式和每个数据分区的元素之间的欧式距离，将每个计算任务分发到分布式系统中的若干节点并行计算，构造出分布式数据集 R ，具体步骤包括：

步骤 5.1、通过 Spark 计算引擎为步骤 4.4 中数据集 Z 的每个数据分区创建一个计算任务，每个计算任务完成的逻辑都是将基准数据序列 s 和该数据分区的每个元素代表的短时间序列 l 计算欧式距离；

步骤 5.2、通过分布式 Spark 计算引擎调度步骤 5.1 中创建的计算任务，将计算任务分发到数据分区所在的计算节点，每个计算任务处理其所在节点的分区数据；

步骤 5.3、创建分布式数据集 R ，数据个数和数据集 Z 相同，其元素为 $\langle key_R, value_R \rangle$ 键值对形式， key_R 为元素编号， $value_R$ 为 Z 的每个元素和基准时间序列之间的计算得出的欧式距离数值；

步骤 6、对步骤 5.3 中的分布式数据集 R 根据欧式距离值进行全排序，取距离最小的前 k 个元素，并返回每个元素的索引，其具体步骤包括：

步骤 6.1、通过 Spark 计算引擎对分布式数据集 R 的每个分区的数据根据元素的 $value$ 进行并行计算排序，取每个分区最小的 k 个元素；

步骤 6.2、将 R 的 n 个分区求得的 $n \times k$ 个元素收集到分布式系统中的一个节点上，进行汇总排序，取最小的 k 个元素，并记录计算得到的元素的索引值；

步骤 7、根据步骤 6 得到的索引取得步骤 4 中数据集 Z 中相应的元素，即得到了海量设备运行历史时间序列数据集中与基准时间序列最相似的 k 个子时间序列。

采用上述技术方案所产生的有益效果在于：本发明提供了一种海量设备运行数据的分布式时间序列模式检索方法，在利用分布式 Spark 计算引擎和分布式文件系统 HDFS 的基础上，对海量设备运行时间序列数据组成的分布式数据集进行重新组织变换，将设备运行数据组成的分布式数据集的每个元素转换成相互独立并且保持时间顺序的短时间序列，通过 Spark 计算引擎可以对每个数据分区创建一个计算任务分发到分布式系统的不同计算节点进行计算。从而可以实现时间序列的并行相似性检索效果，解决单机无法处理的海量设备运行时间序列的相似性检索问题，提高海量设备运行时间序列数据的相似性检索效率。

附图说明

图 1 为本发明实施例提供的一种海量设备运行数据的分布式时间序列模式检索方法的流程图；

图 2 为本发明实施例提供的时间序列检索效果图。

具体实施方式

下面结合附图和实施例，对本发明的具体实施方式作进一步详细描述。以下实施例用于说明本发明，但不用来限制本发明的范围。

高铁轴温数据是典型的时间序列数据，本实施例以某铁路局采集的高铁定子轴温数据为例，采用本发明的面向海量高铁轴温数据基于 Spark 的并行时间序列相似模式检索方法对该时间序列的相似模式进行检索。

高铁定子轴温数据为安装在高铁定子端的轴温传感器采集到的温度数据，本实施例中使用的历史轴温的数据量为 20GB，数据记录条数为 213002710，采样周期为 1s，数据内容包括采集时间和该时刻的轴承温度值，部分数据如表 1 所示。

表 1 部分高铁定子轴温数据

序号	时间	温度值
1	15/7/11/13/4/23	97
2	15/7/11/13/4/24	97
3	15/7/11/13/4/25	97
4	15/7/11/13/4/26	98
5	15/7/11/13/4/27	97
6	15/7/11/13/4/28	98
7	15/7/11/13/4/29	98
8	15/7/11/13/4/30	98
9	15/7/11/13/4/31	98
10	15/7/11/13/4/32	99
11	15/7/11/13/4/33	99
12	15/7/11/13/4/34	100
13	15/7/11/13/4/35	100
14	15/7/11/13/4/36	100
15	15/7/11/13/4/37	99
16	15/7/11/13/4/38	100
17	15/7/11/13/4/39	100
18	15/7/11/13/4/40	100
19	15/7/11/13/4/41	99
20	15/7/11/13/4/42	99
21	15/7/11/13/4/43	99
22	15/7/11/13/4/44	98
23	15/7/11/13/4/45	98

24	15/7/11/13/4/46	98
25	15/7/11/13/4/47	98
26	15/7/11/13/4/48	98
27	15/7/11/13/4/49	98
28	15/7/11/13/4/50	98
29	15/7/11/13/4/51	97
30	15/7/11/13/4/52	98
31	15/7/11/13/4/53	97
32	15/7/11/13/4/54	98
33	15/7/11/13/4/55	97
34	15/7/11/13/4/56	97
35	15/7/11/13/4/57	97
36	15/7/11/13/4/58	97
37	15/7/11/13/4/59	97
38	15/7/11/13/5/0	97
39	15/7/11/13/5/1	97
40	15/7/11/13/5/2	97
41	15/7/11/13/5/3	97
42	15/7/11/13/5/4	98
43	15/7/11/13/5/5	98
44	15/7/11/13/5/6	98
45	15/7/11/13/5/7	98
46	15/7/11/13/5/8	97
47	15/7/11/13/5/9	97
48	15/7/11/13/5/10	97
49	15/7/11/13/5/11	97
50	15/7/11/13/5/12	97
51	15/7/11/13/5/13	97
52	15/7/11/13/5/14	97
53	15/7/11/13/5/15	98
54	15/7/11/13/5/16	98
55	15/7/11/13/5/17	98
56	15/7/11/13/5/18	99
57	15/7/11/13/5/19	98
58	15/7/11/13/5/20	98
59	15/7/11/13/5/21	99
60	15/7/11/13/5/22	99

由于历史轴温数据量大，这里只列出部分定子轴温数据用以说明高铁定子轴温数据的形式。

一种面向海量高铁定子轴温数据基于 Spark 的并行时间序列检索方法，如图 1 所示，包括以下步骤：

步骤 1、设置检索的轴温数据基准时间序列模式 s ，选取第 160 个数据为起始点，长度 m

为 12 的连续数据作为基准时间序列模式，基准时间序列模式 s 的数据为 [103, 103, 105, 105, 110, 159, 144, 127, 116, 116, 117, 113]，设置检索的最相似的模式个数 k 为 2；时间序列模式检索的任务是在海量的轴温历史数据表示的大型时间序列中查找和基准时间序列模式最相似的若干个长度为 m 的子时间序列；

步骤 2、将待检索的高铁定子轴温历史时间序列数据读取进分布式计算节点的内存中，构造分布式数据集 X ，设置并行计算任务个数 n 为 10。具体包括以下步骤：

步骤 2.1、首先将保存在硬盘存储介质中的高铁定子轴温历史数据通过网络上传到分布式文件系统 HDFS 中；

步骤 2.2、使用分布式 Spark 计算引擎读取存储在 HDFS 上的高铁定子轴温数据；

步骤 2.3、在 Spark 读取定子轴温数据的时候设置数据分区个数为 10，Spark 计算引擎根据数据分区的个数将读取到的定子轴温数据切分为 10 个数据块，这些数据分区组成了分布式数据集 RDD 对象 X ， X 的数据分区形式为 $\{partition1, partition2, \dots, partition10\}$ ，其中每个 partition 代表一个数据分区，分区之间具有时间上的顺序性。

本实施例中，以 partition1 的数据说明分区内部的数据形式，分区 partition1 的轴温记录条数为 20001036 条，partition1 的内部数据形式为 $\{x_1, x_2, \dots, x_{20001036}\}$ ， x_n 代表定子轴温的数值，分区的元素按照时间顺序排列，其他分区的形式和 partition1 相同。

步骤 2.4、Spark 计算引擎根据设定的 RDD 对象的数据分区个数，为每个数据分区创建一个计算任务，将计算任务分发到分布式系统中的不同节点，一个任务只处理其对应的数据分区的定子轴温数据；

步骤 3、为步骤 2 中定子轴温数据组成的分布式数据集 X 构建索引，根据时间顺序，对 X 的元素从 0 开始进行编号，数据集 X 的每个分区的数据转换成

$$\{< index, x_1 >, < index + 1, x_2 >, \dots, < index + n - 1, x_n >\}$$

其中， $index$ 为该分区的首个元素相对于整个数据集的偏移量， n 为该分区的数据记录条数，为分布式数据集的每个元素创建一个索引，每个分区的元素索引的创建是在不同节点独立完成的。

本实施例以分区 partition1 为例说明数据的转换逻辑，构建索引之后，分区 partition1 的数据形式转换为

$$\{< 0, x_1 >, < 1, x_2 >, \dots, < 20001035, x_{20001036} >\}$$

其他分区的数据形式和 partition1 相同。

步骤4、根据基准时间序列 s 的长度,构造出11个和步骤3中的 X 相同的分布式数据集 Y_j ,其中 $j \in (1,11)$,将分布式数据集 X 和辅助分布式数据集进行联结,构造出每个元素是一个长度为12的短时间序列的分布式数据集 Z 。

步骤4.1、构造出11个和步骤3中数据集 X 相同的分布式数据集 Y_j ,其中 $j \in (1,11)$;

步骤4.2、对步骤4.1中的11个辅助分布式数据集构建索引,对于 Y_j 来说,对其所有的元素的索引都减去 j ,即转换之后其首元素的索引为 $-j$,以 Y_1 为例,其首元素的索引为-1,其他元素的索引以此类推递增;

步骤4.3、对步骤3中的数据集 X 和11个辅助分布式数据集进行笛卡尔积操作,对进行笛卡尔积操作的任意两个数据集,保留两个数据集中具有相同索引的元素进行联结。两个集合的笛卡尔积数学表达式为

$$A \times B = \{(x, y) | x \in A \wedge y \in B\}$$

其中, A 和 B 表示两个数据集合, x 和 y 分别表示 A 和 B 中德任意元素, (x, y) 表示两个集合进行笛卡尔积计算之后得到的集合中的元素形式。

两个数据集中具有相同索引的两个元素 $\langle key, value1 \rangle$ 和 $\langle key, value2 \rangle$ 进行联结之后的结果为 $\langle key, (value1, value2) \rangle$ 。

本实施例以 X 和 Y_1 的联结为例说明联结的逻辑, X 和 Y_1 中索引为同为10的元素分别为 $\langle 10, 99 \rangle$ 和 $\langle 10, 100 \rangle$,两个元素联结之后的结果为 $\langle 10, (99, 100) \rangle$,其他元素的联结与此类似。

步骤4.4、经过步骤4.3之后,构造出分布式数据集 Z , Z 的每个元素变为

$$\langle key, (value1, value2, \dots, value12) \rangle$$

其中, key 为元素的索引,从0开始编号, $value$ 转换为长度为12的短时间序列。

本实施例以 Z 中第160个元素,即索引为159的元素说明 Z 中数据的形式, Z 中索引 key 为159的元素形式为

$$\langle 159, (103, 103, 105, 105, 110, 159, 144, 127, 116, 116, 117, 113) \rangle$$

步骤5、通过Spark计算引擎调度计算任务,为步骤4.4中数据集 Z 的10个分区创建10个计算任务,将10个计算任务分发到分布式系统的5个节点进行并行计算。具体步骤包括:

步骤 5.1、通过 Spark 计算引擎为步骤 4.4 中数据集 Z 的 10 个数据分区分别创建一个计算任务，每个计算任务完成的逻辑都是将基准数据序列 s 和数据集 Z 的每个分区的长度为 12 的短时间序列元素计算欧式距离；对于两个 k 维向量 $s(x_{11}, x_{12}, \dots, x_{1k})$ 和 $l(x_{21}, x_{22}, \dots, x_{2k})$ 之间的欧式距离为：

$$dist = \sqrt{\sum_1^n (x_{1k} - x_{2k})^2}$$

步骤 5.2、通过分布式 Spark 计算引擎调度步骤 5.1 中创建的计算任务，将计算任务分发到数据分区所在的计算节点，每个计算任务处理其所在节点的分区数据；

步骤 5.3、创建分布式数据集 R ，数据个数和数据集 Z 相同，其元素为 $\langle key_R, value_R \rangle$ 键值对形式， key_R 为元素编号， $value_R$ 为 Z 的每个元素和基准时间序列之间的计算得出的欧式距离数值。

本实施例中，构造的分布式数据集 R 中分区数同样为 10， R 中的数据元素形式如表 2 所示：

表 2 分布式数据集 R 的部分数据形式

索引(key)	欧式距离(value)
0	89.883
1	89.319
2	88.690
3	87.675
⋮	⋮

步骤 6、对步骤 5.3 创建的分布式数据集根据欧式距离的值进行全排序，取值最小的前 2 个元素，得到其索引值。具体步骤包括：

步骤 6.1、通过 Spark 计算引擎为步骤 5.3 中数据集 R 的每个分区创建一个计算任务，分发到 5 个节点计算。计算每个分区元素值最小的 2 个元素，记录其索引值；

步骤 6.2、将步骤 6.1 中计算得到的 2×10 个局部最小元素通过 Spark 计算引擎收集到分布式系统的主节点上，进行汇总排序，计算得到全局最小的 2 个元素，记录其索引值；

步骤 6.3、根据步骤 6.2 中计算得到的 2 个索引值，获取步骤 4.4 中数据集 Z 中相应索引对应的元素值，即为与基准时间序列 s 最相似的 2 个子时间序列。本实施例中根据基准时间序列 s 在轴温历史数据中检索到的最相似的两个子时间序列分别为起始索引为 401 和 95，检索结果如表 3 所示：

表 3 相似模式检索结果

索引	欧式距离	轴温时间序列数值
70	36.318	{105,105,112,117,122,121,178,132,116,113,115,111}
401	38.716	{106,107,107,108,108,119,128,136,112,110,110,111}

检索方法的效果如图 2 所示，其中图中矩形线框标识的数据为定子轴温基准时间序列模式，椭圆形线框标识的数据为在定子轴温历史时间序列中检索出来的最相似的 2 个子时间序列模式。

最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明权利要求所限定的范围。

权 利 要 求 书

1、一种海量设备运行数据的分布式时间序列模式检索方法，其特征在于：包括以下步骤：

步骤 1、设置检索的设备运行数据基准时间序列模式 s ，确定检索模式的长度 m ，设定需要检索的最相似的模式个数为 k ；

所述设备运行数据，采样周期为 T 秒；

所述基准时间序列模式是在设备运行数据中自定义选取的一个连续时间段的数据，长度为 m ；

步骤 2、将待检索的海量历史设备运行时间序列数据读取进分布式系统计算节点的内存中，构造初始分布式数据集 X ，并确定并行计算任务的个数 n ；

步骤 3、为步骤 2 中分布式数据集 X 构建索引，根据时间顺序对数据集 X 中的每个元素从 0 开始编号，其中每个分区的编号任务是在不同节点并行进行计算的；数据集 X 中的每个元素转换为 $\langle key, value \rangle$ 键值对记录形式，其中， key 为索引编号， $value$ 为对应时刻的设备运行时间序列数值；

步骤 4、根据基准时间序列模式的长度构造 $m-1$ 个辅助分布式数据集 Y_j ，其中 $j \in (1, m-1)$ ，将步骤 3 中分布式数据集 X 和辅助分布式数据集进行联结，构造出每个元素是一个长度为 m 的短时间序列的分布式数据集 Z ；

步骤 5、通过 Spark 计算引擎创建计算任务，每个计算任务的逻辑是计算基准时间序列模式和每个数据分区的元素之间的欧式距离，将每个计算任务分发到分布式系统中的若干节点并行计算，构造出分布式数据集 R ；

步骤 6、对步骤 5 中的分布式数据集 R 根据欧式距离值进行全排序，取距离最小的前 k 个元素，并返回每个元素的索引；

步骤 7、根据步骤 6 得到的索引取得步骤 4 中数据集 Z 中相应的元素，即得到了海量设备运行历史时间序列数据集中与基准时间序列最相似的 k 个子时间序列。

2、根据权利要求 1 所述的一种海量设备运行数据的分布式时间序列模式检索方法，其特征在于：所述步骤 2 的具体方法为：

步骤 2.1、将保存在硬盘存储介质中的历史设备运行数据通过网络上传至分布式文件系统 HDFS（Hadoop Distributed File System）中；

步骤 2.2、使用分布式 Spark 计算引擎读取存储在 HDFS 的海量设备运行时间序列数据；

步骤 2.3、设置要创建的分布式数据集的分区个数为 n ，Spark 计算引擎根据设定的分区数将存储在 HDFS 上的设备运行数据切分出 n 个数据块，为每个数据块创建一个分区，构造出分区数为 n 的分布式数据集 RDD（Resilient Distributed Data Sets，即弹性分布式数据集）对

象 X ， X 的每个元素为某一时刻的设备运行数据值，并且 X 维护着分区的顺序及每个分区首个元素在整个数据集中的偏移量；

步骤 2.4、根据分区数确定并行计算任务个数为 n ，Spark 计算引擎为 RDD 对象 X 的每个数据分区创建一个计算任务分发到不同计算节点，任务之间相互独立，每个任务并行处理一个分区的数据。

3、根据权利要求 1 所述的一种海量设备运行数据的分布式时间序列模式检索方法，其特征在于：所述步骤 4 的具体方法为：

步骤 4.1、根据设定的基准时间序列模式的长度 m ，构造出 $m-1$ 个和步骤 3 中数据集 X 相同的辅助分布式数据集 Y_j ，其中 $j \in (1, m-1)$ ；

步骤 4.2、对 $m-1$ 个辅助分布式数据集 Y_j 构建索引，其中 Y_j 的元素从 $-j$ 开始编号；

步骤 4.3、对分布式数据集 X 依次和 $m-1$ 个辅助分布式数据集 Y_j 进行笛卡尔积操作，并将进行笛卡尔积的两个分布式数据集中具有相同索引 key 值的元素进行联结；

步骤 4.4、经过步骤 4.3 之后，从步骤 3 中的数据集 X 构造出分布式数据集 Z ， Z 的元素为 $\langle key_1, (value_1, value_2, \dots, value_m) \rangle$ 形式，其中， key_1 为元素的编号， $(value_1, value_2, \dots, value_m)$ 代表一段以第 key_1 个时刻开始长度为 m 的短时间序列。

4、根据权利要求 3 所述的一种海量设备运行数据的分布式时间序列模式检索方法，其特征在于：所述步骤 5 的具体方法为：

步骤 5.1、通过 Spark 计算引擎为步骤 4.4 中数据集 Z 的每个数据分区创建一个计算任务，每个计算任务完成的逻辑都是将基准数据序列 s 和该数据分区的每个元素代表的短时间序列 l 计算欧式距离；

步骤 5.2、通过分布式 Spark 计算引擎调度步骤 5.1 中创建的计算任务，将计算任务分发到数据分区所在的计算节点，每个计算任务处理其所在节点的分区数据；

步骤 5.3、创建分布式数据集 R ，数据个数和数据集 Z 相同，其元素为 $\langle key_R, value_R \rangle$ 键值对形式， key_R 为元素编号， $value_R$ 为 Z 的每个元素和基准时间序列之间的计算得出的欧式距离数值。

5、根据权利要求 4 所述的一种海量设备运行数据的分布式时间序列模式检索方法，其特征在于：所述步骤 6 的具体方法为：

步骤 6.1、通过 Spark 计算引擎对分布式数据集 R 的每个分区的数据根据元素的 $value$ 进

行并行计算排序，取每个分区最小的 k 个元素；

步骤 6.2、将 R 的 n 个分区求得的 $n \times k$ 个元素收集到分布式系统中的一个节点上，进行汇总排序，取最小的 k 个元素，并记录计算得到的元素的索引值。

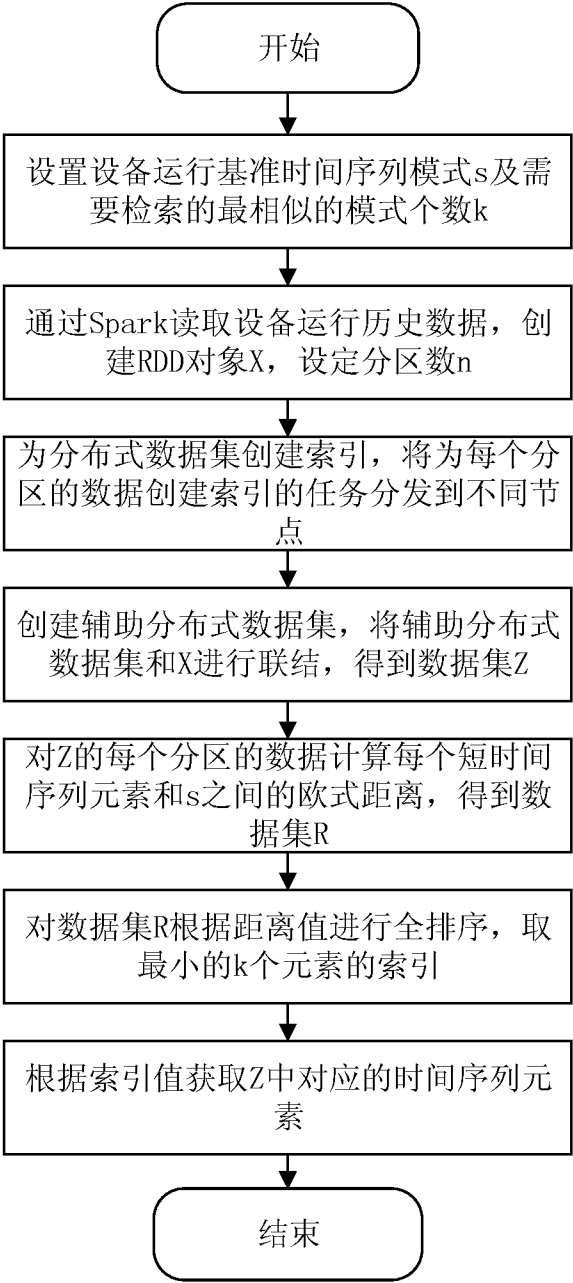


图 1

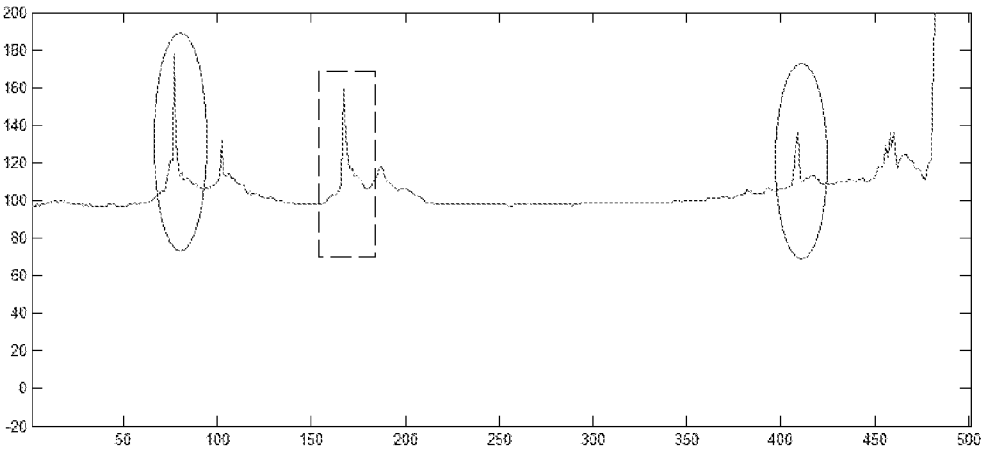


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/077741

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/2458(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)

WPI, EPODOC, CNPAT, CNKI, IEEE: 分布式, 时间序列, 相似, 模式, 检索, 匹配, 笛卡尔积, 连接, 欧式距离, 索引, 聚类, time series, similarity, pattern, euclidean distance, search, index, join, cartesian product, key, value, hadoop

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 108549696 A (ANHUI UNIVERSITY OF TECHNOLOGY) 18 September 2018 (2018-09-18) description, paragraphs [0067]-[0110]	1-5
Y	封朝永 (FENG, Chaoyong). "基于Hadoop的时态信息存储与检索策略的研究 (Research on the Strategy of Temporal Information Storage and Retrieve Based on Hadoop)" 中国优秀硕士学位论文全文数据库(信息科技辑) (Chinese Master's Dissertations Full-text Database (Information Science and Technology)), No. 10, 15 October 2014 (2014-10-15), ISSN: 1674-0246, pp. 31-33 and 39-42	1-5
A	ESLING, Philippe et al. "Time-Series Data Mining" ACM Computing Surveys, Vol. 45, No. 1, 30 November 2012 (2012-11-30), pp. 1-34	1-5
A	CN 107590143 A (BEIJING KINGSOFT CLOUD NETWORK TECHNIQUE CO., LTD. et al.) 16 January 2018 (2018-01-16) entire document	1-5
A	CN 104182460 A (ZHEJIANG UNIVERSITY) 03 December 2014 (2014-12-03) entire document	1-5



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

01 September 2019

Date of mailing of the international search report

10 September 2019

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/CN2019/077741

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2005114331 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 26 May 2005 (2005-05-26) entire document	1-5

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/077741

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	108549696	A	18 September 2018	None	
CN	107590143	A	16 January 2018	None	
CN	104182460	A	03 December 2014	None	
US	2005114331	A1	26 May 2005	None	

国际检索报告

国际申请号

PCT/CN2019/077741

A. 主题的分类

G06F 16/2458(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))

WPI, EPDOC, CNPAT, CNKI, IEEE: 分布式, 时间序列, 相似, 模式, 检索, 匹配, 笛卡尔积, 连接, 欧式距离, 索引, 聚类, time series, similarity, pattern, euclidean distance, search, index, join, cartesian product, key, value, hadoop

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
Y	CN 108549696 A (安徽工业大学) 2018年 9月 18日 (2018 - 09 - 18) 说明书第[0067]-[0110]段	1-5
Y	封朝永. “基于Hadoop的时态信息存储与检索策略的研究” 中国优秀硕士学位论文全文数据库(信息科技辑), 第10期, 2014年 10月 15日 (2014 - 10 - 15), ISSN: 1674-0246, 第31-33页, 第39-42页	1-5
A	ESLING, Philippe 等. “Time-Series Data Mining” ACM Computing Surveys, 第45卷, 第1期, 2012年 11月 30日 (2012 - 11 - 30), 第1-34页	1-5
A	CN 107590143 A (北京金山云网络技术有限公司 等) 2018年 1月 16日 (2018 - 01 - 16) 全文	1-5
A	CN 104182460 A (浙江大学) 2014年 12月 3日 (2014 - 12 - 03) 全文	1-5

☒ 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

2019年 9月 1日

国际检索报告邮寄日期

2019年 9月 10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

曾璇

电话号码 86-(10)-53961373

国际检索报告

国际申请号

PCT/CN2019/077741

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
A	US 2005114331 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2005年 5月 26日 (2005 - 05 - 26) 全文	1-5

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/077741

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	108549696	A	2018年 9月 18日	无	
CN	107590143	A	2018年 1月 16日	无	
CN	104182460	A	2014年 12月 3日	无	
US	2005114331	A1	2005年 5月 26日	无	