

(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
2017 年 8 月 31 日 (31.08.2017)



(10) 国际公布号
WO 2017/143908 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2017/073135
- (22) 国际申请日: 2017 年 2 月 9 日 (09.02.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610096728.8 2016 年 2 月 22 日 (22.02.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 代斌 (DAI, Bin); 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)。 杨旭 (YANG, Xu); 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)。 姜晓燕 (JIANG, Xiaoyan); 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)。 蔡宁 (CAI, Ning); 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)。 王少萌 (WANG, Shaomeng); 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: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

[见续页]

(54) Title: ASSOCIATION ANALYSIS METHOD AND DEVICE

(54) 发明名称: 关联分析方法和装置

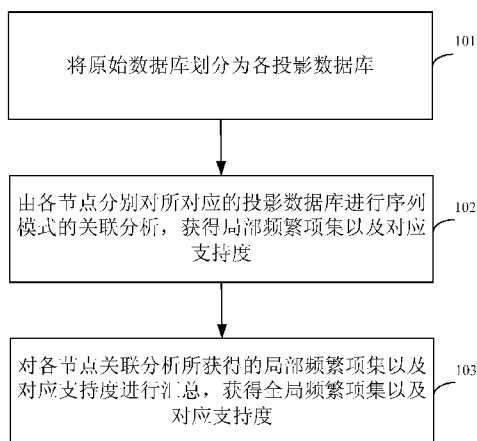


图 1

- 101 DIVIDING AN ORIGINAL DATABASE INTO PROJECTION DATABASES
- 102 NODES RESPECTIVELY PERFORMING AN ASSOCIATION ANALYSIS OF A SEQUENTIAL PATTERN ON CORRESPONDING PROJECTION DATABASES TO OBTAIN LOCAL FREQUENT ITEM SETS AND CORRESPONDING SUPPORT DEGREES
- 103 SUMMARIZING THE LOCAL FREQUENT ITEM SETS AND THE CORRESPONDING SUPPORT DEGREES OBTAINED BY THE NODES THROUGH ASSOCIATION ANALYSIS SO AS TO OBTAIN A GLOBAL FREQUENT ITEM SET AND A CORRESPONDING SUPPORT DEGREE

(57) Abstract: Provided are an association analysis method and device. The method comprises: dividing an original database into projection databases not contributing a support degree of a frequent item set to one another; nodes respectively performing an association analysis of a sequential pattern on corresponding projection databases; and then, summarizing local frequent item sets and corresponding support degrees obtained by the nodes. Since the established projection databases do not contribute a support degree of a frequent item set to one another, different nodes can respectively perform association mining, comprising the step of pruning, on different projection databases; and local frequent item sets with a relatively small data amount are obtained by the nodes so as to avoid the situation in the prior art of needing to transmit local candidate frequent item sets with a relatively large data amount obtained by nodes without the step of pruning, such that the transmission overhead is saved and the efficiency is improved.

(57) 摘要: 本发明提供了关联分析方法和装置, 通过将原始数据库划分为相互之间不贡献频繁项集支持度的各投影数据库, 由各节点分别对对应的投影数据库进行序列模式的关联分析, 然后对各节点获得的局部频繁项集以及对应支持度进行汇总。由于所建立的投影数据库相互之间不贡献频繁项集的支持度, 可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘, 各节点所获得的为数据量较小的局部频繁项集, 避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况, 从而节省了传输开销, 提高了效率。



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。

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)。

(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,

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

关联分析方法和装置

本申请要求 2016 年 02 月 22 日递交的申请号为 201610096728.8、发明名称为“关联分析方法和装置”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。

5 技术领域

本发明涉及计算机技术，尤其涉及一种关联分析方法和装置。

背景技术

关联分析是进行数据挖掘中的一种分析技术，主要用于根据大量数据发现项目之间的
10 的关联性。关联分析的一个典型应用实例便是购物篮分析：基于购物数据进行关联分析，发现顾客放入购物篮中的不同商品之间的关联性。进而由这种关联分析所获得的关联性体现出顾客的购买习惯，通过了解这些购买习惯可以有利于零售商制定营销策略。

序列模式关联分析不同于其他模式下的关联分析，序列模式还考虑了项目发生的时间，从而使得各项目之间有一定的顺序性。针对序列模式的关联分析，其所进行分析的
15 对象往往是超大规模的数据库，因此，运算量很大不适宜采用单机进行数据处理。现有技术中通常采用对数据库中的数据进行简单分片，将每个分片数据在各节点进行单独的关联分析，获得候选频繁项集及其支持度，然后进行合并获得各候选频繁项集的全局支持度，进而依据预设的筛选条件进行剪枝后获得全局频繁项集。

但是，由于在进行关联分析生成候选频繁项集的过程中存在数据膨胀，导致候选频繁
20 项集的数据量是分片数据的数据量的指数倍，因此，在对各节点生成的候选频繁项集及其支持度进行汇总以便执行合并操作时，传输数据量过大，导致执行效率较低。

发明内容

本发明提供一种关联分析方法和装置，用于解决现有技术中进行并行关联分析时由
25 于汇总数据时数据传输量过大导致执行效率较低的技术问题。

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，提供一种关联分析方法，包括：

将原始数据库划分为各投影数据库，所述各投影数据库相互之间不贡献频繁项集的支持度；

30 由各节点分别对所对应的投影数据库进行序列模式的关联分析，获得局部频繁项集

以及对应支持度；

对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总，获得全局频繁项集以及对应支持度。

第二方面，提供一种关联分析装置，包括：

5 划分模块，用于将原始数据库划分为各投影数据库，所述各投影数据库相互之间不贡献频繁项集的支持度；

分析模块，用于由各节点分别对所述投影数据库进行序列模式的关联分析，获得局部频繁项集以及对应支持度；

10 汇总模块，用于对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总，获得全局频繁项集以及对应支持度。

本发明实施例提供的关联分析方法和装置，通过将原始数据库划分为相互之间不贡献支持度的各投影数据库，由各节点分别对所对应的投影数据库进行序列模式的关联分析，然后对各节点获得的局部频繁项集以及对应支持度进行汇总。由于所建立的投影数据库相互之间不贡献频繁项集的支持度，可以由不同节点分别对不同的投影数据库进行
15 包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够
20 更明显易懂，以下特举本发明的具体实施方式。

附图说明

通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的
25 限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 为本发明实施例一提供的一种关联分析方法的流程示意图；

图 2 为本发明实施例二提供的一种关联分析方法的流程示意图；

图 3 为关联分析方法执行过程的示意图；

图 4 为本发明实施例所提供的一种关联分析装置的结构示意图；

30 图 5 为实施例三提供的另一种关联分析装置的结构示意图；

图 6 为实施例四提供一种关联分析装置的结构示意图。

具体实施方式

下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的
5 示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例
所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

在待进行关联分析的数据库中，通常记载了各个事务，事务又包括了各个项目，每一个项目包括至少一个元素。其中，一个元素用于指示一个操作对象，项目用于指示由
10 同一用户同时进行操作的操作对象，而一个用户在不同时刻所进行操作的操作对象可以用一个事务标识。由于序列模式下的关联分析需要考虑时序性，因此，每一个事务所包括的各个项目是具有一定顺序性的，这种具有顺序的各个项目可以称之为一个序列，因此，也可以说事务是由一个序列进行表示的。

比如数据库中的一个事务为序列 `abc,abc,ac,d,cf`，可以用于表示一个用户分别在第一天买了商品 `a`、`b` 和 `c`，第二天又买了商品 `a`、`b` 和 `c`，第三天买了商品 `a` 和 `c`。针对每一
15 天买的总商品叫项目，每件商品叫元素。

发明人针对现有技术中的关联分析方法进行分析，发现现有技术中在对各节点生成的候选频繁项集及其支持度进行汇总以便执行合并操作时，传输数据量过大，主要是由于各节点未在本地执行剪枝的步骤。发明人在此基础上，进行了进一步分析，现有技术
20 中对原始数据仅进行了简单分片，各个分片数据对于某个频繁项级都存在贡献支持度的可能，从而无法在本地执行剪枝的步骤，因此，需要将原始数据划分为相互之间不存在支持度的数据库才能够在节点本地进行剪枝，进而避免传输候选频繁项集及其支持度。基于这一思路，发明人提出了本发明所提供的关联分析方法。

为了便于理解本发明所提供的关联分析方法，在描述具体实施例之前，对实施例中
25 所涉及的技术术语进行解释：

序列模式关联分析是指：给定一个数据库，其中，数据库包括了各个事务，每一个事务由一个序列表示，每个序列由相互之间具有一定顺序性的项目组成。序列模式挖掘就是在给定一个支持度阈值的基础上，找出所有满足在数据库中的出现频次不低于该支持度阈值的子序列，将这些子序列作为频繁项集，从数据库中找出这些频繁项集的操作
30 便是序列模式关联分析。

项集是指：数据库针对同一事务中所出现的项目有序排列所构成的集合。

项集的长度是指：用于表示该项集的序列的长度，数值上等于项集所包括的项目个数。

支持度是指：项集在数据库中出现的频次，若一个事务中包含该项集，则记为一次，
5 从而支持度等于数据库中包含某一项集的事务个数。

频繁项集是指：数据库中所有出现频次不小于支持度阈值的项集。

投影数据库是指：针对原始数据库进行投影操作所获得的数据库，具体来说 α 的投影数据库为 S 中所有以 α 为前缀的序列相对于 α 的后缀。

此处简要解释了技术术语的含义，以上技术术语会在后续结合具体实施例进行进一步地解释。
10

下面结合附图对本发明实施例提供的关联分析方法和装置进行详细描述。

实施例一

图 1 为本发明实施例一提供的一种关联分析方法的流程示意图，如图 1 所示，包括：
步骤 101、将原始数据库划分为各投影数据库。

15 其中，各投影数据库相互之间不贡献频繁项集的支持度。其中，支持度是指在数据库中包含某一项集的事务个数。各投影数据库相互之间不贡献频繁项集的支持度，也就是说基于一投影数据库进行序列模式挖掘所获得的频繁项集未出现在另一投影数据库中，从而另一投影数据库不会增加频繁项集的支持度，因此说另一投影数据库不贡献一投影数据库的频繁项集的支持度。

20 具体地，首先，获得至少两个初始项集，其中初始项集是对原始数据库进行关联分析所获得的项集，用于构建投影数据库，并且至少两个初始项集中的任意两初始项集之间不存在相互包含关系。具体可以通过针对原始数据库可以采用关联分析算法进行计算，获得前述的至少两项集，这两项集可以是相同长度也可以是不同长度，可以是执行过剪枝步骤获得的频繁项集也可以是未执行过剪枝步骤获得的候选频繁项集，本实施例中对此不做限定。
25

进而，将各初始项集作为前缀，建立各前缀的投影数据库。具体可以通过将初始项集作为前缀，在原始数据库的各事务中查询前缀所对应的后缀，进而将各事务的后缀进行汇总，形成前缀的投影数据库。其中，后缀是通过在每一条事务中查询该前缀首次出现的位置之后的序列，若前缀的最后一个元素与所查找到的序列的第一个元素的时序相同，则将 “_” 和所查找到的序列作为后缀，否则，直接将所查找到的序列作为后缀。
30

由于初始项集两两之间的不存在包含关系，所以据此所建立的投影数据库相互之间不贡献频繁项集的支持度，可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

需要说明的是，这里定义的包含关系是前缀的包含关系，也就说一个初始项集是否为另一初始项集的前缀，若是，则存在包含关系，例如：

针对 $\langle a, b \rangle$ 和 $\langle a, c, b \rangle$ ， $\langle a, b \rangle$ 不是 $\langle a, c, b \rangle$ 的前缀，因此 $\langle a, b \rangle$ 和 $\langle a, c, b \rangle$ 两者之间不存在包含关系；

针对 $\langle a, b \rangle$ 和 $\langle a \rangle$ ， $\langle a \rangle$ 是 $\langle a, b \rangle$ 的前缀，因此 $\langle a, b \rangle$ 和 $\langle a \rangle$ 属于包含关系。

作为一种可能的实现方式，可以计算原始数据库的 K _频繁项集，也就是项集中包含 K 个项目的频繁项集，其中 K 为正整数，且 $1 \leq K < N$ ， N 为原始数据库中所包含的元素数，将 K _频繁项集作为初始项集。

作为另一种可能的实现方式，还可以计算原始数据库的 K _频繁项集之后，根据预设的支持度阈值，对 K _频繁项集进行筛选，保留支持度大于支持度阈值的 K _频繁项集，将筛选后的 K _频繁项集作为初始项集。经过筛选步骤之后，合理减小了后续进行处理的数据量，减轻了节点构建投影数据库以及基于投影数据库进行关联分析的运算压力，同时，也减少了后续对节点关联分析所获得的局部频繁项集进行汇总时的数据传输总量。

步骤 102、由各节点分别对所对应的投影数据库进行序列模式的关联分析，获得局部频繁项集以及对应支持度。

具体的，为每一投影数据库分配节点，例如：可以为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点。然后由各节点采用预设关联分析算法并行进行序列模式的关联分析，如广义序贯模式(Generalized Sequential Pattern, GSP)算法，具体来说，各节点对投影数据库执行扫描、合并和剪枝的步骤，获得频繁项集以及对应支持度。每个节点仅能够获得原始数据库的各频繁项集中的一部分频繁项集，为了与原始数据库的全部频繁项集进行区分，将每个节点所获得的频繁项集称为局部频繁项集，将原始数据库的全部频繁项集称为全局频繁项集。另外需要说明的是，这里所说的节点运行在一个单机上，单机可以是物理机也可以是虚拟机本实施例中对此不做限定。

通过为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点，避免出现有些节点负载过重而另一些节点空闲的情况出现，从而更加合理和高效地利用现有

节点，加快关联分析速度，提高关联分析的效率。

步骤 103、对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总，获得全局频繁项集以及对应支持度。

具体地，将各节点关联分析所获得的局部频繁项集以及对应支持度汇总到一个文件既可，从而该文件中记录的为全局频繁项集以及对应支持度，而无需执行合并和剪枝的步骤。

本实施例中，通过将原始数据库划分为相互之间不贡献支持度的各投影数据库，由各节点分别对所对应的投影数据库进行序列模式的关联分析，然后对各节点获得的局部频繁项集以及对应支持度进行汇总。由于所建立的投影数据库相互之间不贡献频繁项集的支持度，可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

实施例二

图 2 为本发明实施例二提供的一种关联分析方法的流程示意图，本实施例所提供的关联分析方法可以由软件执行，该软件可以运行在大数据计算服务（Open Data Processing Service, 简称 ODPS）平台上，本实施例所提供的方法可以运行在映射规约（MapReduce）软件框架下由多个节点执行，在 MapReduce 软件框架下包括一个主节点（master），和多个从节点（workers），主节点可以对各从节点分配映射（Map）作业（用来把一组键值对映射成一组新的键值对）或者规约（Reduce）作业（用来保证所有映射的键值对中的每一个共享相同的键组），分配到 Map 作业的从节点又可以称为 Map 节点，分配到 Reduce 作业的从节点又可以称为 Reduce 节点。每一个节点可以运行在一个物理机或者虚拟机上，本实施例中对此不做限定。如图 2 所示，方法包括：

步骤 201、主节点对从节点进行调度以使从节点进行数据预处理。

具体地，主节点将数据预处理作为一项 MapReduce 任务，调度从节点执行该任务，从而针对原始数据进行数据序列化操作。具体通过根据映射表将原始数据中所记载的各用户执行操作的不同的操作对象分别用数字或者字母序列进行标识，并针对同一用户执行操作的操作对象按照操作时间进行排序。

步骤 202、从节点在主节点调度下基于原始数据库 D 计算 K_频繁项集，将每一个

K_频繁项集分配至一个 Map 节点进行处理。

需要说明的是, $1 \leq K < N$, N 为原始数据库中所包含的元素数。

步骤 203、各 Map 节点根据每一个 K_频繁项集遍历原始数据库 D 中的一个分片数据, 获得每一个 K_频繁项集的投影数据库 D' 的一部分。

5 具体地, 每一个 Map 节点均根据各 K_频繁项集将原始数据库 D 的一个分片数据中的数据划分为前缀和后缀两部分, 由全部 Map 节点所获得的对应某一个 K_频繁项集的后缀部分所构成的投影数据库 D' 即为该 K_频繁项集的投影数据库 D'。具体来说, 在各事务中, 将该 K_频繁项集作为前缀, 查询该前缀首次出现的位置之后的序列, 若前缀的最后一个元素与所查找到的序列的第一个元素的时序相同, 则将“_”和所查找到的序列作为后缀, 否则, 直接将所查找到的序列作为后缀。构建由各事务中的后缀所构成的该 K_频繁项集的投影数据库 D'。

Map 节点的个数可以为多个, 各 Map 节点获取到原始数据库 D 中的一个分片数据, 这里的分片是简单的数据分片, 每一个 Map 节点将各 K_频繁项集作为键, 遍历分片数据中的各事务, 获得各个键或者说各个 K_频繁项集对应的键值, 将所获得的键值输出至该 K_频繁项集对应的 Reduce 节点, 从而该 K_频繁项集对应的 Reduce 节点从全部 Map 节点所接收到的数据构成了该 K_频繁项集的投影数据库 D'。例如: 若存在 m 个 K_频繁项集, 3 个分片数据时, Map 节点 1 基于 m 个 K_频繁项集将原始数据库 D 的第一个分片数据划分为前缀和后缀两部分, Map 节点 2 基于 m 个 K_频繁项集将原始数据库 D 的第二个分片数据划分为前缀和后缀两部分, Map 节点 3 基于 m 个 K_频繁项集将原始数据库 D 的第三个分片数据划分为前缀和后缀两部分。

需要说明的是, 投影数据库 D' 中“_”用于表示前缀自身, 且该前缀为后缀中第一个项目的元素, 该第一个项目除了该前缀外还有其他元素。

对于相同前缀的各事务, 可以看出其生成的频繁项集也具有相同的前缀, 因此将各事务中同一前缀首次出现的位置之后的序列合并为投影数据库, 不同投影数据库所关联分析获得的频繁项集不会出现重复的情况, 因而也就互不贡献支持度。从而不同的 Reduce 节点可以针对不同的投影数据库独立的进行关联分析挖掘, 包括对候选频繁项集剪枝获得频繁项集的过程也在本地 Reduce 节点, 避免了汇总候选频繁项集及其本地支持度。

可见, 本实施例中所提供的方法不同于目前不同的 Reduce 节点关联分析获得的候选频繁项集会出现重复的情况, 因而不需要汇总 Reduce 节点所关联分析获得的候选频繁项集其本地支持度之后, 才能够进行合并和剪枝, 最终获得全局频繁项集。因此, 本实施

例中的方法，能够有效避免汇总候选频繁项集及其本地支持度，而各 Reduce 节点候选频繁项集的往往是该 Reduce 节点的分片数据的数据量的指数倍，因此，本实施例中的方法极大减少了进行数据传输的数据量。

步骤 204、各 Reduce 节点从全部 Map 节点接收所对应的 K_频繁项集的投影数据库 D' 的一部分，获得所对应的 K_频繁项集的投影数据库 D'，对投影数据库 D' 进行并
5 行关联分析处理。

具体的，MapReduce 软件框架包括多个 Reduce 节点，每一个 Reduce 节点对应一个 K_频繁项集。每一个 Reduce 节点从全部 Map 节点接收所对应的 K_频繁项集的投影数据库 D' 的一部分，从而获得所对应的 K_频繁项集的投影数据库 D'，进而对该 K_频繁
10 项集的投影数据库 D' 进行关联分析处理，获得频繁项集及其支持度。

例如：Reduce 节点 1 基于投影数据库 D'_1 进行关联分析处理，Reduce 节点 2 基于投影数据库 D'_2 进行关联分析处理，……Reduce 节点 m 基于投影数据库 D'_m 进行关联分析处理。

其中，各 Reduce 节点可以采用 GSP 算法进行关联分析处理从而获得前述的频繁项
15 集及其支持度，也可以采用其他关联分析算法而不会影响本实施例所提供的方法的使用效果，本实施例中 GSP 算法仅作为示例说明本实施例所提供的方法。GSP 算法是通过扫描投影数据库 D' 得到长度为 i 的序列作为初始序列，然后根据长度为 i 的初始序列，经过合并和剪枝的操作，生成长度为 i+1 的序列，并将生成的序列作为新的初始序列，重复迭代执行扫描、合并和剪枝的操作，直至不再生成新的序列，用所获得的序列表示候
20 选频繁项集。基于投影数据库 D' 计算候选频繁项集的支持度，然后根据预设的筛选条件进行筛选，获得频繁项集。

需要说明的是，i 的初始取值应当等于作为前缀的频繁项集的序列长度。

具体可以采用如下所示的 GSP 算法伪代码：

Algorithm GSP (S)

```

25 1 C1 ← init-pass(S);
2 F1 ← {<{f}> | f ∈ C1, f.count/n ≥ minsup};
3 for(k=2; Fk-1 ≠ ∅; k++) do
4   Ck ← candidate-gen-SPM(Fk-1);
5   for each data sequence s ∈ S do
30 6     for each candidate c ∈ Ck do

```

```

7         if c is contained in s then
8             c.count++
9         end
10    end
5    11     $F_k \leftarrow \{c \in C_k \mid c.count/n \geq \text{minsup}\}$ 
12 end
13 return  $\cup_k F_k$ ;

```

其中, candidate-gen-SPM(F_{k-1})的算法流程如下:

1、合并: 对所生成的 $k-1$ _候选频繁项集 F_{k-1} 进行合并, 生成 k _候选频繁项集 F_k 。

10 具体地, 当 $k=2$ 时, 对于两个 1 _候选项集 F_1 的序列 s_1 和 s_2 , 需要将 s_2 的项目要以 s_1 的项目中的一部分和以一个单独的项目两种方式合并到 s_1 。即合并 $\langle a \rangle$ 和 $\langle b \rangle$, 生成的候选频繁项级有 $\langle (a, b) \rangle$ 、 $\langle a, a \rangle$ 、 $\langle a, b \rangle$ 、 $\langle b, a \rangle$ 和 $\langle b, b \rangle$ 。

需要说明的是, $\langle (a, b) \rangle$ 表示 a, b 同时发生, $\langle a, b \rangle$ 表示先发生 a , 后发生 b 。

15 当 k 取大于 2 的正整数时, 对于两个序列 s_1 和 s_2 , 如果将 s_1 的第 1 个项目去掉后得到的余串和将 s_2 的最后一个项目去掉后得到的余串相同, 则可以将 s_1 和 s_2 合并。所得的候选序列是将 s_2 的最后一个项目添加到 s_1 末尾, 这里针对两种不同情况有两种添加方式:

如果 s_2 最后一个项目是一个单独的元素, 则这个项目将以一个单独项目的形式加到 s_1 的末尾, 否则, s_2 最后一个项目将作为 s_1 的最后一个项目中的一个部分合并入 s_1 。

2、剪枝: 对合并所获得的 k _候选频繁项集 F_k 进行剪枝, 获得 k _频繁项集 F_k 。

20 具体地, 剪枝是指如果一个 k _候选频繁项集 F_k 的任何一个子集是非频繁的, 则这个 k _候选频繁项集 F_k 将被去除。

Reduce 节点基于前述扫描、合并和剪枝的步骤以及筛选的步骤, 获得关联分析处理结果, 即频繁项集及其支持度。

步骤 205、各 Reduce 节点对关联分析处理结果输出。

25 具体的, 各节点可以直接输出关联分析处理结果, 还可以在输出之前, 根据映射表对结果进行反序列化处理, 从而输出反序列化处理后的关联分析处理结果。

步骤 206、对各 Reduce 节点输出的关联分析处理结果进行汇总。

30 具体的, 可以主节点调度从节点直接合并各 Reduce 节点输出的结果, 汇总为一个文件既可。由于各 K _频繁项集的投影数据库 D' 相互之间不贡献支持度, 可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘, 各节点所输出的关联分析

处理结果为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

为了清楚说明本实施例，本实施例提供了一个具体的实例以对本实施例所提供的方法进行详细说明。

5 例如：

图 3 为关联分析方法执行过程的示意图，如图 3 所示，针对表 1 中的原始数据进行序列化。在表 1 中，原始数据的每一行对应一个用户，列向量从左至右依次为用户标识和操作对象以及对该操作对象执行操作的操作时间。

表 1 原始数据的记录表

甲用户	苹果梨桃	2 月 1 日	苹果梨桃	2 月 2 日	苹果桃	2 月 3 日
乙用户	梨桃	2 月 3 日	苹果	2 月 4 日	苹果梨	2 月 5 日
丙用户	苹果	2 月 6 日	梨	2 月 7 日	香蕉	2 月 8 日

10 表 1 中的原始数据记录到了甲、乙、丙用户分别于不同日期执行了三次购买操作，根据如下映射表：

苹果→a

梨→b

桃→c

15 香蕉→d

执行数据序列化操作，获得原始数据库 D。在原始数据库 D 中所记载的内容如下所示：

abc, abc, ac;

bc,a, ab;

20 a,b,d。

需要说明的是，每一行代表一个事务，在每一个事务中，用逗号分隔各个项目。

基于前述原始数据库 D，进行序列模式的关联分析，计算 1_候选频繁项集及其支持度，如下表所示。

表 2 1_候选频繁项集及其支持度

候选频繁项集	支持度
<a>	3
	3
<c>	2
<d>	1

若预先设置了以支持度 2 作为最小支持度阈值，也就是说支持度小于 2 的候选频繁项集将会被过滤掉，从而获得 1_频繁项集<a>，，<c>。

在每一个 Map 节点中构建各 1_频繁项集的投影数据库的一部分，Reduce 节点从每个 Map 节点接收对应 1_频繁项集的投影数据库一部分，汇总获得该 1_频繁项集的投影数据库，进行关联分析处理。

将 1_频繁项集<a>作为前缀 a，构建前缀 a 的投影数据 D'_a ，所构建的前缀 a 的投影数据库 D'_a 如下所示：

_bc,abc,ac

ab

b,d。

将 1_频繁项集作为前缀 b，构建前缀 b 的投影数据 D'_b ，所构建的前缀 b 对应的投影数据库 D'_b 如下所示：

_c,abc,ac

_c,a,ab

d。

将 1_频繁项集<c>作为前缀 c，构建前缀 c 的投影数据 D'_c ，所构建的前缀 c 对应的投影数据库 D'_c 如下所示：

abc,ac

a,ab

在前面的步骤中已获得 1_频繁项集<a>，，<c>。根据 1_频繁项集所获得的 2_候选频繁项集如下所示：

<(a, b)>，<(a, c)>，<(b, c)>，<a, a>，<a, b>，<a, c>，<b, a>，<b, b>，<b, c>，<c, a>，<c, b>，<c, c>。

需要说明的是，<(a, b)>表示 a, b 同时发生，<a, b>表示先发生 a，后发生 b。

数据库 D 已被划分为前缀 a 的投影数据库 D'_a ，前缀 b 的投影数据库 D'_b ，前缀 c 的投影数据库 D'_c ，并分别由不同 Reduce 节点进行关联分析处理，如在 Reduce 节点 1 中，对于<(a, b)>，<(a, c)>，<a, a>，<a, b>，<a, c>只需要基于投影数据库 D'_a 计算支持度。

表 3 Reduce 节点 1 的支持度计算结果

频繁项集	支持度
<(a, b)>	2
<(a, c)>	1
<a, a>	2
<a, b>	3
<a, c>	1

同理，Reduce 节点 2 对于<(b, c)>, <b, a>, <b, b>, <b, c>只需要基于投影数据库 D'_b 计算支持度。

表 4 Reduce 节点 2 的支持度计算结果

频繁项集	支持度
<(b, c)>	2
<b, a>	2
<b, b>	2
<b, c>	1

Reduce 节点 3 对于<c, a>, <c, b>, <c, c>只需要基于投影数据库 D'_c 计算支持度。

表 5 Reduce 节点 3 的支持度计算结果

频繁项集	支持度
<c, a>	2
<c, b>	2
<c, c>	1

可以针对各个 Reduce 节点设置相同的支持度阈值作为筛选条件，当频繁项集的支持度大于该支持度阈值时，筛选通过并保留，否则筛选掉而不进行保留。当支持度阈值为 1 时，各 Reduce 节点分别保留如下频繁项集：

表 6 Reduce 节点 1 的筛选结果

频繁项集	支持度
<(a, b)>	2
<a, a>	2
<a, b>	3

表 7 Reduce 节点 2 的筛选结果

频繁项集	支持度
<(b, c)>	2
<b, a>	2
<b, b>	2

表 8 Reduce 节点 3 的筛选结果

频繁项集	支持度
<c, a>	2
<c, b>	2

各个 Reduce 节点对筛选后的频繁项集，根据映射表进行反序列化处理，进而各个

Reduce 节点输出反序列化处理后的关联分析处理结果，表 9 为各 Reduce 节点输出的反序列化处理后的关联分析处理结果示意。

表 9 各 Reduce 节点输出的结果

	频繁项集	支持度
Reduce 节点 1	<(苹果, 梨)>	2
	<苹果, 苹果>	3
	<苹果, 梨>	3
Reduce 节点 2	<(梨, 桃)>	2
	<梨, 苹果>	2
	<梨, 梨>	2
Reduce 节点 3	<桃, 苹果>	2
	<桃, 梨>	2

对各 Reduce 节点输出的结果进行汇总，汇总获得的文件内容如下所示：

表 10 汇总结果

频繁项集	支持度
<(苹果, 梨)>	2
<苹果, 苹果>	3
<苹果, 梨>	3
<(梨, 桃)>	2
<梨, 苹果>	2
<梨, 梨>	2
<桃, 苹果>	2
<桃, 梨>	2

本实施例中，通过对原始数据库进行关联分析计算，获得至少两个初始项集之后，将各初始项集作为前缀，建立各前缀的投影数据库，由各节点分别对所对应的投影数据库进行序列模式的关联分析，然后对各节点获得的局部频繁项集以及对应支持度进行汇总。由于至少两个初始项集之间不存在相互包含关系，所以据此所建立的投影数据库相互之间不贡献频繁项集的支持度，可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

实施例三

图 4 为本发明实施例所提供的一种关联分析装置的结构示意图，如图 4 所示，包括：划分模块 42、分析模块 43 和汇总模块 44。

划分模块 42, 用于将原始数据库划分为各投影数据库, 所述各投影数据库相互之间不贡献频繁项集的支持度。

分析模块 43, 用于由各节点分别对所述投影数据库进行序列模式的关联分析, 获得局部频繁项集以及对应支持度。

5 具体地, 分析模块 43 具体用于采用预设关联分析算法, 由节点对所述投影数据库执行扫描、合并和剪枝的步骤, 获得局部频繁项集以及对应支持度。其中, 关联分析算法包括 GSP 算法。

汇总模块 44, 用于对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总, 获得全局频繁项集以及对应支持度。

10 进一步, 本实施例还提供了一种关联分析装置的结构示意图, 图 5 为实施例三提供的另一种关联分析装置的结构示意图, 如图 5 所示, 划分模块 42, 包括: 获得单元 421、建立单元 422。

获得单元 421, 用于对原始数据库进行序列模式的关联分析, 获得至少两个初始项集。

15 其中, 所述至少两个初始项集之间不存在相互包含关系。

建立单元 422, 用于将各初始项集作为前缀, 建立各前缀的投影数据库。

进一步, 获得单元 421, 包括: 计算子单元 4211、确定子单元 4212 筛选子单元 4213。

计算子单元 4211, 用于计算原始数据库的 k _频繁项集。

其中 k 为正整数, 且 $1 \leq k < N$, N 为原始数据库中所包含的元素数。

20 确定子单元 4212, 用于将所述 k _频繁项集作为所述初始项集。

筛选子单元 4213, 用于根据预设的支持度阈值, 对 k _频繁项集进行筛选, 保留支持度大于所述支持度阈值的 k _频繁项集。

进一步, 建立单元 422, 包括: 查询子单元 4221 和建立子单元 4222。

25 查询子单元 4221, 用于将所述初始项集作为前缀, 在所述原始数据库的各事务中查询所述前缀所对应的后缀;

建立子单元 4222, 用于将各事务的后缀进行汇总形成所述前缀的投影数据库。

本实施例中, 通过对原始数据库进行关联分析计算, 获得至少两个初始项集之后, 将各初始项集作为前缀, 建立各前缀的投影数据库, 由各节点分别对所对应的投影数据库进行序列模式的关联分析。由于至少两个初始项集之间不存在相互包含关系, 所以据
30 此所建立的投影数据库相互之间不贡献频繁项集的支持度, 可以由不同节点分别对不同

的投影数据库进行包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。

5 实施例四

图 6 为实施例四提供的一种关联分析装置的结构示意图，在上一实施例的基础上，本实施例中的装置，还包括：分配模块 61。

分配模块 61，用于为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点。

10 进一步，关联分析装置还包括：序列化模块 62 和反序列化模块 63。

序列化模块 62，用于根据映射表，对原始数据进行序列化获得原始数据库。

反序列化模块 63，用于对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总之前，对所述局部频繁项集进行反序列化；或者，获得全局频繁项集以及对应支持度之后，对所述全局频繁项集进行反序列化。

15 本实施例中，通过对原始数据库进行关联分析计算，获得至少两个初始项集之后，将各初始项集作为前缀，建立各前缀的投影数据库，由各节点分别对所对应的投影数据库进行序列模式的关联分析。由于至少两个初始项集之间不存在相互包含关系，所以据此所建立的投影数据库相互之间不贡献频繁项集的支持度，可以由不同节点分别对不同的投影数据库进行包括剪枝步骤在内的关联挖掘，各节点所获得的为数据量较小的局部
20 频繁项集，避免现有技术中需要传输各节点未经过剪枝步骤所获得的数据量较大的局部候选频繁项集的情况，从而节省了传输开销，提高了效率。另外，针对各投影数据库可能具有不同的规模的特点，为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点，避免出现有些节点较为空闲，而另外一些节点过载的情况发生从而进一步提高关联分析的效率

25 本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一计算机可读取存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各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
30

依然可以对前述各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各实施例技术方案的范围。

权 利 要 求 书

1.一种关联分析方法，其特征在于，包括：

将原始数据库划分为各投影数据库，所述各投影数据库相互之间不贡献频繁项集的支持度；

5 由各节点分别对所对应的投影数据库进行序列模式的关联分析，获得局部频繁项集以及对应支持度；

对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总，获得全局频繁项集以及对应支持度。

2.根据权利要求1所述的关联分析方法，其特征在于，所述将原始数据库划分为各
10 投影数据库，包括：

对所述原始数据库进行序列模式的关联分析，获得至少两个初始项集；其中，各初始项集之间不存在相互包含关系；

将各初始项集作为前缀，建立各前缀的投影数据库。

3.根据权利要求2所述的关联分析方法，其特征在于，所述对原始数据库进行关联
15 分析计算，获得至少两个初始项集，包括：

计算原始数据库的 K _频繁项集；其中 K 为正整数，且 $1 \leq K < N$ ， N 为原始数据库中所包含的元素数；

将所述 K _频繁项集作为所述初始项集。

4.根据权利要求3所述的关联分析方法，其特征在于，所述将 K _频繁项集作为所述
20 初始项集之前，还包括：

根据预设的支持度阈值，对 K _频繁项集进行筛选，保留支持度大于所述支持度阈值的 K _频繁项集。

5.根据权利要求2所述的关联分析方法，其特征在于，所述将各初始项集作为前缀，建立各前缀的投影数据库，包括：

25 将所述初始项集作为前缀，在所述原始数据库的各事务中查询所述前缀所对应的后缀；

将各事务的后缀进行汇总形成所述前缀的投影数据库。

6.根据权利要求1所述的关联分析方法，其特征在于，所述由各节点分别对所对应的投影数据库进行序列模式的关联分析之前，包括：

30 为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点。

7.根据权利要求1所述的关联分析方法,其特征在于,所述由各节点分别对所对应的投影数据库进行序列模式的关联分析,获得局部频繁项集以及对应支持度,包括:

采用预设关联分析算法,由所述节点对所述投影数据库执行扫描、合并和剪枝的步骤,获得局部频繁项集以及对应支持度。

5 8.根据权利要求7所述的关联分析方法,其特征在于,所述关联分析算法包括 GSP 算法。

9.根据权利要求1-8任一项所述的关联分析方法,其特征在于,所述对原始数据库进行关联分析计算,获得至少两个初始项集之前,还包括:

根据映射表,对原始数据进行序列化获得原始数据库。

10 10.根据权利要求9所述的关联分析方法,其特征在于,

所述对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总之前,对所述局部频繁项集进行反序列化;

或者,获得全局频繁项集以及对应支持度之后,对所述全局频繁项集进行反序列化。

11.一种关联分析装置,其特征在于,包括:

15 划分模块,用于将原始数据库划分为各投影数据库,所述各投影数据库相互之间不贡献频繁项集的支持度;

分析模块,用于由各节点分别对所述投影数据库进行序列模式的关联分析,获得局部频繁项集以及对应支持度;

20 汇总模块,用于对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总,获得全局频繁项集以及对应支持度。

12.根据权利要求11所述的关联分析装置,其特征在于,所述划分模块,包括:

获得单元,用于对原始数据库进行序列模式的关联分析,获得至少两个初始项集;其中,各初始项集之间不存在相互包含关系;

建立单元,用于将各初始项集作为前缀,建立各前缀的投影数据库。

25 13.根据权利要求12所述的关联分析装置,其特征在于,所述获得单元,包括:

计算子单元,用于计算原始数据库的 K _频繁项集;其中 K 为正整数,且 $1 \leq K < N$, N 为原始数据库中所包含的元素数;

确定子单元,用于将所述 K _频繁项集作为所述初始项集。

14.根据权利要求13所述的关联分析装置,其特征在于,所述获得单元,还包括:

30 筛选子单元,用于根据预设的支持度阈值,对 K _频繁项集进行筛选,保留支持度

大于所述支持度阈值的 K 频繁项集。

15. 根据权利要求 12 所述的关联分析装置，其特征在于，所述建立单元，包括：

查询子单元，用于将所述初始项集作为前缀，在所述原始数据库的各事务中查询所述前缀所对应的后缀；

5 建立子单元，用于将各事务的后缀进行汇总形成所述前缀的投影数据库。

16. 根据权利要求 11 所述的关联分析装置，其特征在于，所述装置，还包括：

分配模块，用于为各投影数据库分配负载能力与所述投影数据库的数据量相匹配的节点。

17. 根据权利要求 11 所述的关联分析装置，其特征在于，

10 所述分析模块，具体用于采用预设关联分析算法，由节点对所述投影数据库执行扫描、合并和剪枝的步骤，获得局部频繁项集以及对应支持度。

18. 根据权利要求 17 所述的关联分析装置，其特征在于，所述关联分析算法包括 GSP 算法。

19. 根据权利要求 11-18 任一项所述的关联分析装置，其特征在于，所述装置，还包

15 括：

序列化模块，用于根据映射表，对原始数据进行序列化获得原始数据库。

20. 根据权利要求 19 所述的关联分析装置，其特征在于，所述装置，还包括：

反序列化模块，用于对各节点关联分析所获得的局部频繁项集以及对应支持度进行汇总之前，对所述局部频繁项集进行反序列化；或者，获得全局频繁项集以及对应支持度之后，对所述全局频繁项集进行反序列化。

20

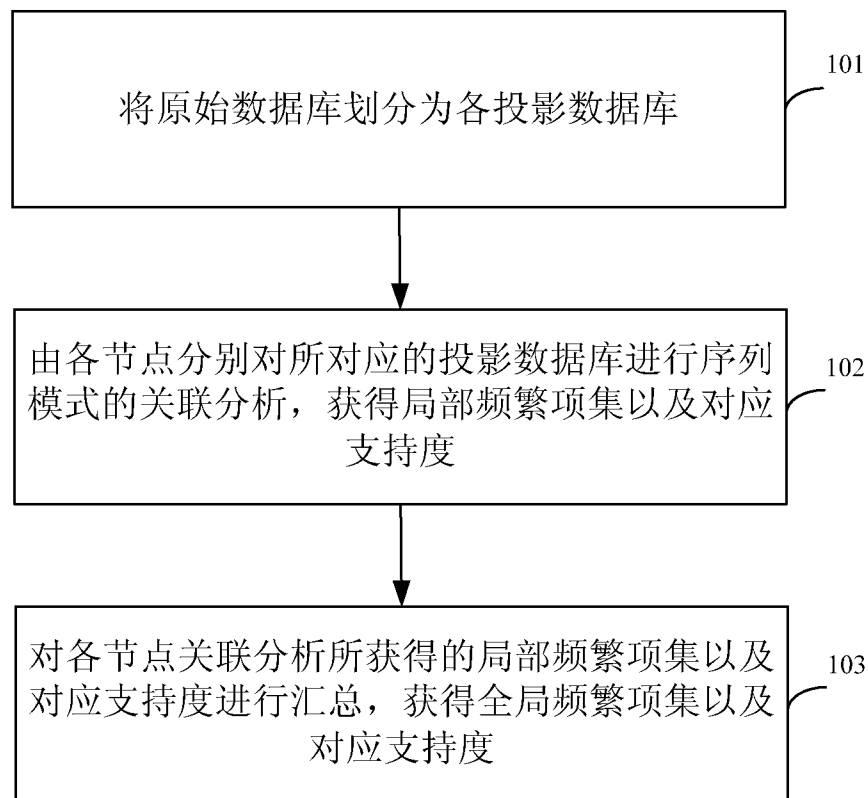


图 1

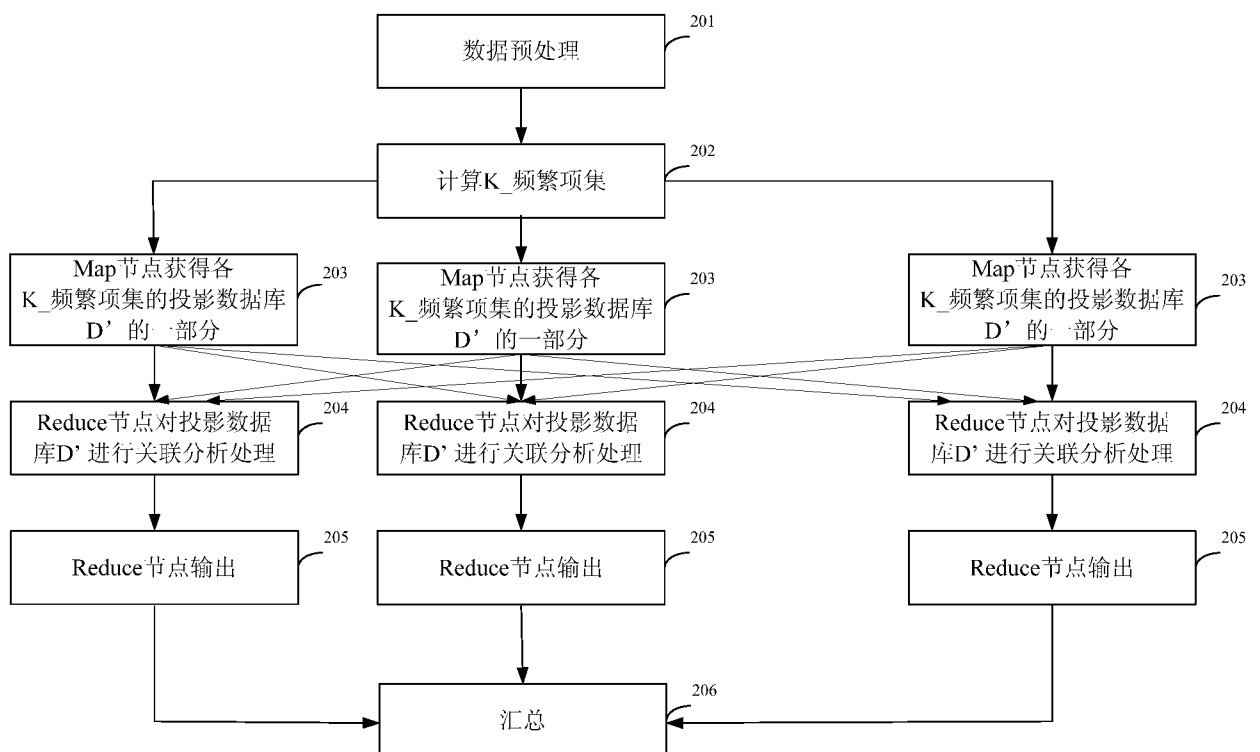
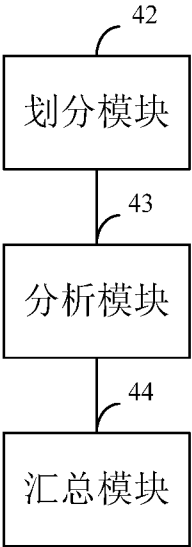
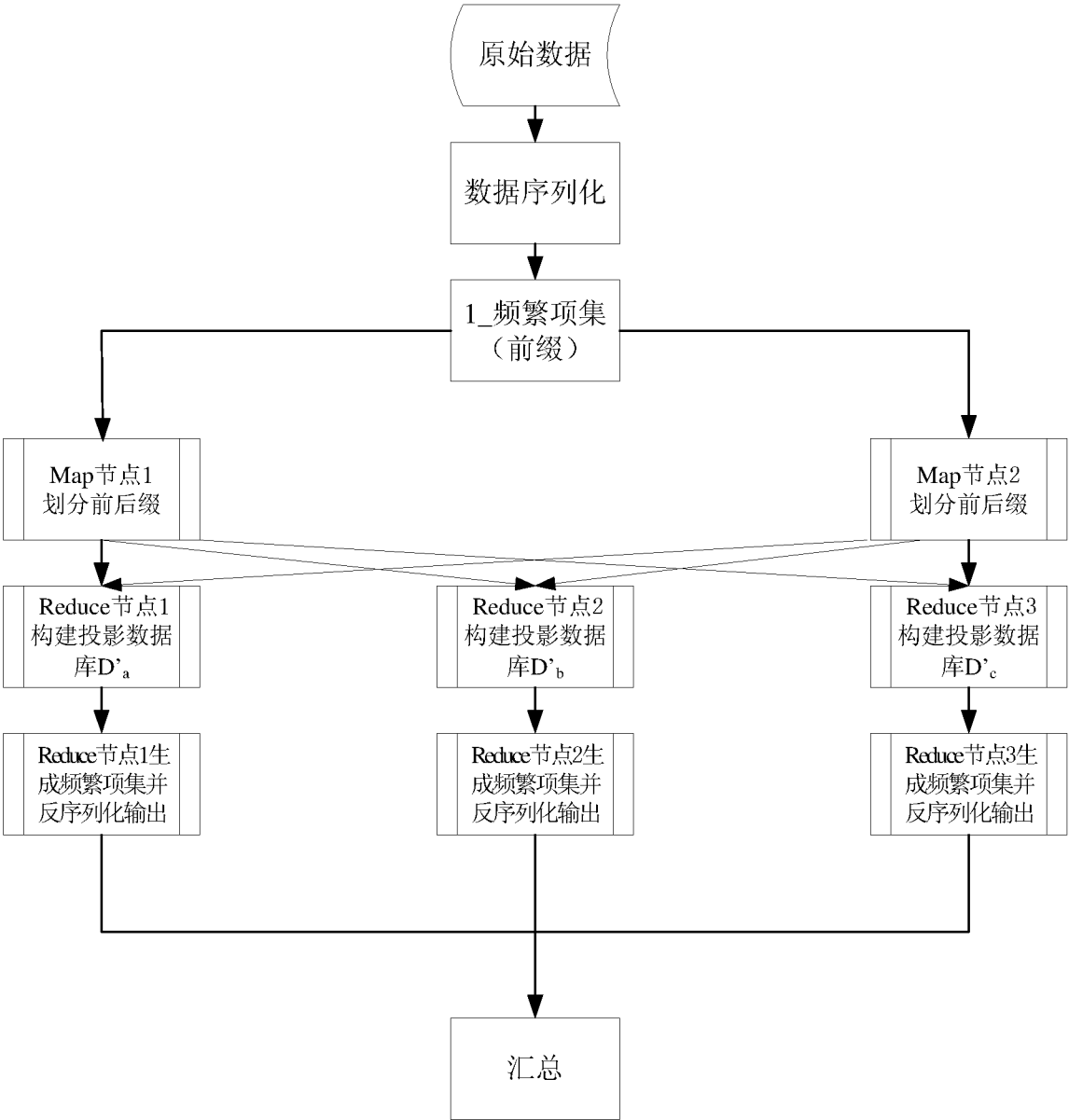


图 2



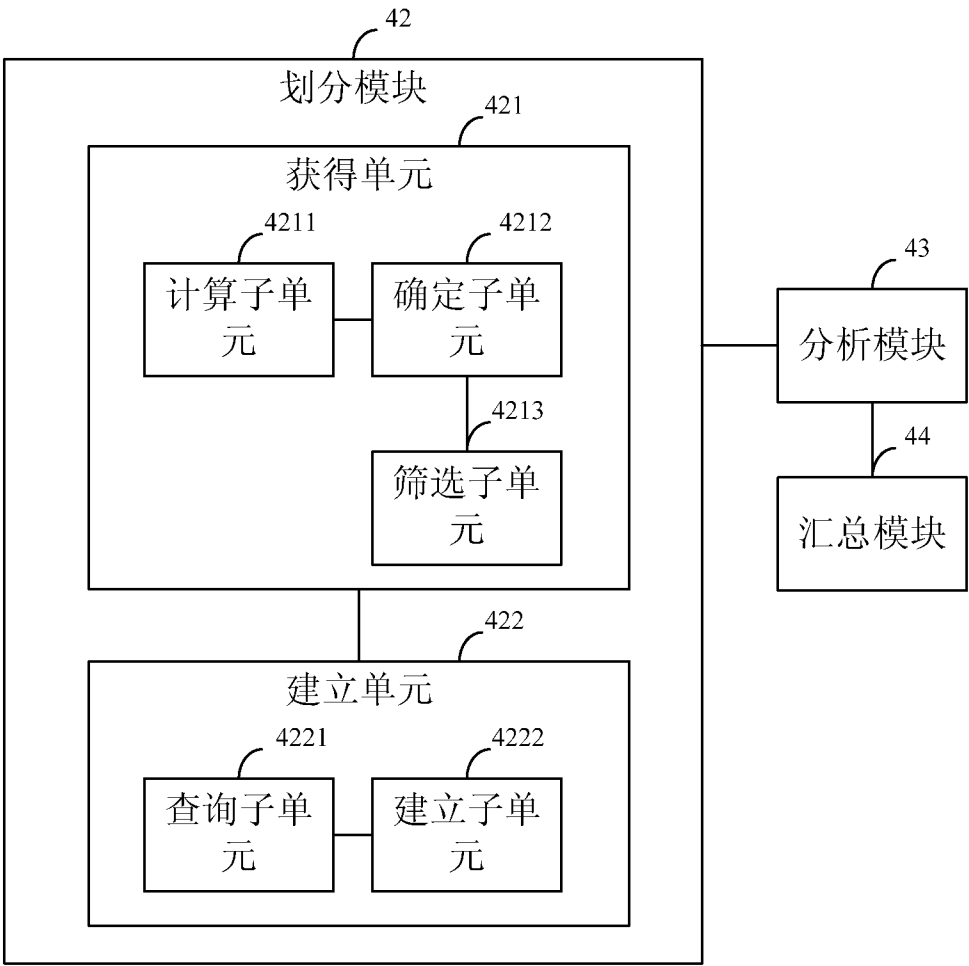


图 5

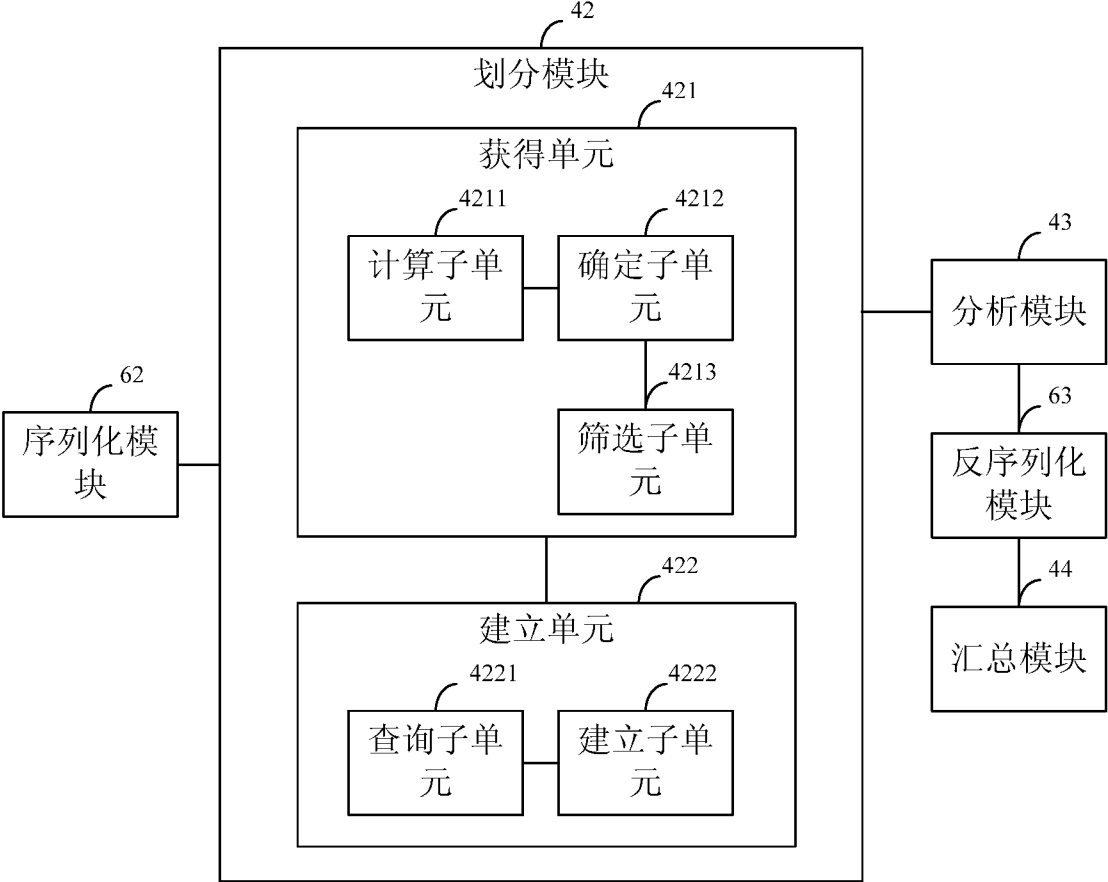


图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/073135

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

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)

CNABS, SIPOABS, DWPI, CNKI: database, projective, frequency item set, support, node, global, local, association rule

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 104834751 A (CHENGDU YICHEN DEXUN SCIENCE & TECHNOLOGY CO., LTD.) 12 August 2015 (12.08.2015) claims 1-3	1-20
Y	CN 102541934 A (BEIJING SAFE CODE TECHNOLOGY CO., LTD.) 04 July 2012 (04.07.2012) claims 4-6	1-20
A	US 2011289590 A1 (NOKIA CORPORATION) 24 November 2011 (24.11.2011) 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 March 2017	Date of mailing of the international search report 26 April 2017
Name and mailing address of the ISA 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 KANG, Jian Telephone No. (86-10) 62411639

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/073135

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104834751 A	12 August 2015	None	
CN 102541934 A	04 July 2012	None	
US 2011289590 A1	24 November 2011	US 8397304 B2	12 March 2013
		WO 2009127771 A1	22 October 2009

国际检索报告

国际申请号

PCT/CN2017/073135

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) 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)) CNABS, SIPOABS, DWPI, CNKI: 数据库, 投影, 频繁项集, 支持度, 节点, 全局, 局部, 关联规则, database, projective, frequency itemset, support, node, global, local, association rule														
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>CN 104834751 A (成都艺辰德迅科技有限公司) 2015年 8月 12日 (2015 - 08 - 12) 权利要求1-3</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 102541934 A (北京安码科技有限公司) 2012年 7月 4日 (2012 - 07 - 04) 权利要求4-6</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2011289590 A1 (NOKIA CORPORATION) 2011年 11月 24日 (2011 - 11 - 24) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	CN 104834751 A (成都艺辰德迅科技有限公司) 2015年 8月 12日 (2015 - 08 - 12) 权利要求1-3	1-20	Y	CN 102541934 A (北京安码科技有限公司) 2012年 7月 4日 (2012 - 07 - 04) 权利要求4-6	1-20	A	US 2011289590 A1 (NOKIA CORPORATION) 2011年 11月 24日 (2011 - 11 - 24) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求												
Y	CN 104834751 A (成都艺辰德迅科技有限公司) 2015年 8月 12日 (2015 - 08 - 12) 权利要求1-3	1-20												
Y	CN 102541934 A (北京安码科技有限公司) 2012年 7月 4日 (2012 - 07 - 04) 权利要求4-6	1-20												
A	US 2011289590 A1 (NOKIA CORPORATION) 2011年 11月 24日 (2011 - 11 - 24) 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 2017年 3月 20日		国际检索报告邮寄日期 2017年 4月 26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		受权官员 康健 电话号码 (86-10) 62411639												

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/073135

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	104834751	A	2015年 8月 12日	无			
CN	102541934	A	2012年 7月 4日	无			
US	2011289590	A1	2011年 11月 24日	US	8397304	B2	2013年 3月 12日
				WO	2009127771	A1	2009年 10月 22日