



(12) 发明专利申请

(10) 申请公布号 CN 104778258 A

(43) 申请公布日 2015.07.15

(21) 申请号 201510187447.9

(22) 申请日 2015.04.21

(71) 申请人 华中科技大学

地址 430074 湖北省武汉市洪山区珞喻路
1037 号

(72) 发明人 王非 潘鑫侨

(74) 专利代理机构 武汉东喻专利代理事务所
(普通合伙) 42224

代理人 宋业斌

(51) Int. Cl.

G06F 17/30(2006.01)

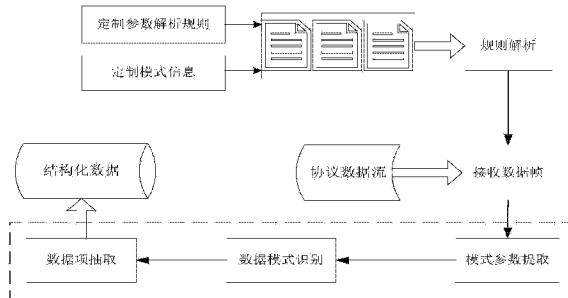
权利要求书4页 说明书10页 附图4页

(54) 发明名称

一种面向协议数据流的数据抽取方法

(57) 摘要

本发明公开了一种面向协议数据流的数据抽取方法，属于数据仓库领域。本发明根据工业领域中数据帧的结构特点，给出了一种通用的面向协议数据流的数据抽取机制，具体包括三个步骤：(1) 提取描述性信息，获取抽取数据项所需要的解析参数；(2) 利用解析参数，确定数据帧中数据域的模式信息，主要包括数据域的结构、格式和类型；(3) 根据解析参数和数据域的模式信息，实现对数据项的抽取，并转换成结构化的数据保存。本发明可以对各种类型的协议数据帧的数据实现有效而准确地抽取，更能保证数据抽取的扩展性、通用性以及灵活性，即在通信协议改变的情况下本发明也能适用。



1. 一种面向协议数据流的数据抽取方法,其特征在于,包括:

步骤 1 建立针对工业通信协议的数据抽取规则,所述数据抽取规则包括参数解析规则和模式信息规则,分别保存于参数解析规则文件和模式信息规则文件,其中,所述参数解析规则文件用于描述数据帧的类型与结构,所述模式信息规则文件用于描述数据转换与处理规则;

步骤 2 读取所述参数解析规则文件,将其中所描述的不同解析节点实例化成具体的对象,生成由实例对象构成的数据解析树;读取所述模式信息规则文件,生成数据模式映射表;

步骤 3 从协议数据流中接收数据帧,并利用所述数据解析树从所述数据帧中提取出关于数据项内容的描述性信息,以获取抽取数据项所需要的解析参数;

步骤 4 利用所述解析参数以及所述数据模式映射表,确定所述数据帧中数据域的模式信息,其中,所述模式信息包括所述数据域的结构、格式和类型;

步骤 5 根据所述解析参数以及所述模式信息,利用有限状态机实现对所述数据帧中数据项的抽取,并将抽取的数据项转换成结构化的数据保存。

2. 如权利要求 1 所述的方法,其特征在于,所述参数解析规则分为原子参数解析结构以及组合参数解析结构,其中,所述原子参数解析结构包含六个主要属性,采用六元组来描述: $In = \langle Ad, Be, L, D, Rn, DT \rangle$, 其中 In 表示原子参数解析结构, Ad 表示位置分布类型, Be 表示语义单元起始地址, L 表示原子语义单元的长度, D 表示与源数据实现逻辑与的操作数, Rn 表示需要右偏移位数, DT 表示目标数据格式;所述组合参数解析结构包含两个主要属性,采用二元组来描述: $InS = \langle G, Lin \rangle$, 其中 InS 表示组合参数解析结构, G 表示组合规则,即通过 G 将多个原子信息单元组合成有实际意义的信息单元, Lin 表示参数解析集,包含了多个原子参数解析结构 In 。

3. 如权利要求 1 所述的方法,其特征在于,所述解析节点选择方式分为基于特征字选择方式和基于令牌选择方式,其中:

所述基于特征字选择方式包含一个哈希映射表用于特征字的快速查询,该哈希映射表包含了所有解析节点特征字的哈希表,可用二元组 $(key, value)$ 表示,其中 key 对应解析节点的特征字, $value$ 为 key 对应的解析节点的名称;

所述基于令牌选择方式包含一个模式信息链表,所述模式信息链表中的每个解析节点由三部分组成:逻辑算子,通过所述逻辑算子以及抽取参数集中的对应参数来判断是否选择当前的解析节点;抽取参数集,为与当前逻辑算子相对应的相关参数;节点名称,表示当前解析节点对应的解析节点名称。

4. 如权利要求 1-3 中任一项所述的方法,其特征在于,所述步骤 2 包括以下子步骤:

在步骤 201 中,获取解析节点的配置信息,读取所述参数解析规则文件,依次将其中的配置信息转换为 XML 格式对应的参数解析规则 DOM 对象;

在步骤 202 中,根据所述参数解析规则 DOM 对象判断当前解析节点的参数解析类型,如果只有一个原子解析参数,则执行步骤 203,否则执行步骤 204;

在步骤 203 中,读取所述参数解析规则 DOM 对象中的参数并相应赋予原子参数解析结构的各属性,然后执行步骤 205;

在步骤 204 中,从所述参数解析规则 DOM 对象中,按照所述步骤 203 的方式读取多个原

子参数解析结构，并构建组合参数解析结构，将读取的多个原子参数解析结构赋予参数解析集，并进一步读取组合规则以及参数解析集；

在步骤 205 中，根据所述参数解析规则 DOM 对象中抽取参数标签下各子标签内容是否为空来判断当前解析节点的解析节点选择方式，如果不为空则执行步骤 206，否则执行步骤 207；

在步骤 206 中，创建基于哈希表的解析节点映射表，将解析节点名称作为 Key 值，用在内存中建立的对应解析节点对象的引用作为 Value 值，构建 (Key, Value) 对插入到所述解析节点哈希映射表中，然后执行步骤 208；

在步骤 207 中，依次读取解析节点名称、逻辑算子以及抽取参数集，并将其按所述参数解析规则文件中的顺序插入到模式信息链表中；

在步骤 208 中，判断所述参数解析规则文件是否读取完成，是则执行步骤 209，否则执行步骤 201；

在步骤 209 中，读取所述模式信息规则文件，依次读取其中各个数据项的模式信息；

在步骤 210 中，读取所述模式信息中的各个元组信息，并采用链式结构将其存储于内存中；

在步骤 211 中，以模式标识作为 Key 值，对应的模式信息链表入口地址作为 Value 值建立基于哈希表的数据模式映射表；

在步骤 212 中，判断所述模式信息规则文件是否读取完毕，是则执行所述步骤 3，否则继续执行步骤 209。

5. 如权利要求 1-3 中任一项所述的方法，其特征在于，所述步骤 3 包括以下子步骤：

在步骤 301 中，接收待解析的数据帧；

在步骤 302 中，根据解析树根节点的名称到解析节点对象映射表中获取根解析节点对象；

在步骤 303 中，执行当前解析节点的解析函数，判断当前参数解析类型，如果是原子参数类型则执行步骤 304，否则执行步骤 307；

在步骤 304 中，读取解析节点对象中配置的抽取参数信息，并根据起始地址和长度定位到当前参数的数据帧区域，获取数据项内容；

在步骤 305 中，用逻辑与操作数与所述步骤 304 中获取的数据项内容做逻辑与操作，将结果再按右偏移位数向右偏移，获取参数在数据区域的有效位；

在步骤 306 中，根据目标数据格式将从数据帧中获取的数据转换成目标数据，然后执行步骤 312；

在步骤 307 中，遍历模式信息链表，依次读取抽取参数的信息；

在步骤 308 中，根据读取的抽取参数中的起始地址和长度定位到当前参数的数据帧区域并获取数据项内容，用逻辑与操作数与获取的数据项内容做逻辑与操作，将结果再按右偏移位数向右偏移，获取参数在数据区域的有效位，根据目标数据格式将从数据帧中获取的数据转换成目标数据；

在步骤 309 中，将所述步骤 308 抽取的参数存储到一个临时的抽取参数链表；

在步骤 310 中，判断是否到达模式信息链表尾，是则执行 311，否则执行 307；

在步骤 311 中，根据组合规则以及已获取的相关参数，得到组合参数数据；

在步骤 312 中,根据参数名称,将获取的参数值输出到共享参数列表中;

在步骤 313 中,判断当前节点的选择方式,如果是基于特征字的选择方式,则执行步骤 314,否则执行步骤 315;

在步骤 314 中,根据所述步骤 312 获取的参数值,从解析节点映射表中获取对应解析节点的名称,然后执行步骤 316;

在步骤 315 中,根据获取的参数值、抽取参数以及逻辑算子,依次遍历模式信息链表,直到得到节点名称;

在步骤 316 中,通过节点名称,获取下一节点的解析节点对象;

在步骤 317 中,判断当前是否到达解析树的叶子节点,是则执行步骤 4,否则执行步骤 303。

6. 如权利要求 1-3 中任一项所述的方法,其特征在于,所述步骤 5 包括以下子步骤:

在步骤 501 中,根据所述步骤 3 获得的解析参数,获取数据项的抽取参数;

在步骤 502 中,从起始偏移地址处开始读取数据帧,定位数据域的初始地址,并将数据模式偏移到首地址处;

在步骤 503 中,根据数据模式的偏移地址获取当前的输入事件以及元组提取参数;

在步骤 504 中,判断当前所处状态以及输入事件,如果当前判断条件为 S_{B_m} ,则执行步骤 505,如果当前判断条件为 S_{B_v} ,则执行步骤 506,如果当前判断条件为 S_{M_v} ,则执行步骤 507,如果当前判断条件为 S_{V_q} ,则执行步骤 508,如果当前判断条件为 S_{V_t} ,则执行步骤 509,如果当前判断条件为 S_{Q_t} ,则执行步骤 510,如果当前判断条件为 S_{V_e} ,则执行步骤 511,如果当前判断条件为 S_{Q_e} ,则执行步骤 512,如果当前判断条件为 S_{T_e} ,则执行步骤 513,如果当前判断条件为 S_{T_b} ,则执行步骤 514,如果当前判断条件为 S_{Q_b} ,则执行步骤 515,如果当前判断条件为 S_{V_b} ,则执行步骤 516,其中,有限状态机的转移状态包括六个状态分别为 B、M、V、Q、T 和 E,其中 B 表示起始状态,表示数据域数据抽取的开始状态或者是数据域中某个数据记录抽取的开始状态,E 表示结束状态,表示当前数据域的数据抽取过程全部完成,M、V、Q、T 状态分别表示提取完数据标识、数据元素值、数据质量以及数据时间后所处的状态;有限状态机的输入事件包括六个输入事件 b、m、v、q、t 和 e,其中 b 事件表示抽取完记录的最后一个元组数据,事件 m、v、q、t 分别表示下一阶段的实际抽取的数据为数据标识、数据元素值、数据质量以及数据时间,e 事件表示抽取完成后记录的最后一个元组数据;

在步骤 505 中,通过数据标识的操作函数和函数参数集,从指定偏移处提取出数据项中的数据标识,数据项循环个数减 1,有限状态机状态转移到状态 M,并执行步骤 517;

在步骤 506 中,根据解析参数中的数据项地址和当前已抽取数据项个数,得到数据记录地址,补齐当前数据记录的数据地址,通过数据值的提取操作函数和函数参数集,从指定偏移处提取出数据项中的数据值,数据项循环个数减 1,有限状态机状态转移到状态 V,并执行步骤 517;

在步骤 507 中,通过数据值的提取操作函数和函数参数集,从指定偏移处提取出数据项中的数据值,有限状态机状态转移到状态 V,并执行步骤 517;

在步骤 508 中,通过数据质量的提取操作函数和函数参数集,从指定偏移处提取出数据项中的数据质量,有限状态机状态转移到状态 Q,并执行步骤 517;

在步骤 509 中,从解析参数中获取到当前数据项的数据质量,补齐当前数据记录的数

据质量,通过数据时间的提取操作函数和函数参数集,从指定偏移处提取出数据项中的数据时间,有限状态机状态转移到状态 T,并执行步骤 517;

在步骤 510 中,通过数据时间的提取操作函数和函数参数集,从指定偏移处提取出数据项中的数据时间,有限状态机状态转移到状态,并执行步骤 517;

在步骤 511 中,从解析参数中获取到当前数据项的数据质量,补齐当前数据记录的数据质量,从解析参数获取到该数据帧的数据时间,补齐当前数据记录的数据时间,有限状态机状态转移到终止状态 E,结束流程;

在步骤 512 中,从解析参数获取到该数据帧的数据时间,补齐当前数据记录的数据时间,有限状态机状态转移到终止状态 E,结束流程;

在步骤 513 中,有限状态机状态转移到终止状态 E,结束流程;

在步骤 514 中,有限状态机状态转移到开始状态 B,并执行步骤 517;

在步骤 515 中,从解析参数获取到该数据帧的数据时间,补齐当前数据记录的数据时间,有限状态机状态转移到开始状态 B,并执行步骤 517;

在步骤 516 中,从解析参数中获取到当前数据项的数据质量,补齐当前数据记录的数据质量,从解析参数获取到该数据帧的数据时间,补齐当前数据记录的数据时间,有限状态机状态转移到开始状态 B,并执行步骤 517;

在步骤 517 中,判断当前是否为数据项的最后一个元组,如果是则执行步骤 518,否则执行步骤 519;

在步骤 518 中,判断当前数据项个数是否为 0,如果是则执行步骤 520,否则执行步骤 521;

在步骤 519 中,数据模式偏移到下一元组属性地址处,并执行步骤 503;

在步骤 520 中,生成结束事件 e,执行步骤 504;

在步骤 521 中,数据模式偏移到首地址处,生成开始事件 b,并执行步骤 504。

一种面向协议数据流的数据抽取方法

技术领域

[0001] 本发明属于数据仓库技术领域,更具体地,涉及一种面向协议数据流的数据抽取方法。

背景技术

[0002] 近年来,随着物联网、云计算、大数据等技术的快速发展,相关的工业设备向着智能化的方向迈进,使得工业信息化的发展进程不断地加快。通过将全球的工业系统与先进计算、数据分析工具、低成本的传感设备和更高联网水平的高度融合,将重构全球工业,提高生产效率,工业的创新和变革正在展开。互联网技术和工业技术的深度融合将深刻地改变人们的生活方式,让世界更快速、更安全、更清洁并且更经济,必将由此引发全球范围内的再一次的技术革命,美国著名公司 GE 将此次技术革命称之为“工业互联网革命”,即是由工业互联网技术而引发的生产力革命。而工业互联网技术能广泛应用的基础即要解决能将海量分布的智能工业设备中的实时数据快速高效地集成到数据仓库中。

[0003] 智能工业设备的智能性即在于设备的行为能够通过软件程序快速灵活地定制设计,例如改变交互过程,协议数据参数或者是应用层的数据通信协议,并且要求数据采集系统要能够同样快速及时地响应这种变化。传统的数据采集技术通过二次开发的方式,也能适应这种变化,但是往往开发周期很长,无法从根本上满足及时快速响应的要求。具体来讲,在对于工业互联网系统更加智能化和信息化的系统而言,传统的数据采集技术在通用性、扩展性以及灵活性方面存在着明显的不足。

[0004] 智能设备对数据采集系统的挑战最根本地在于智能设备能够更灵活地选择各种应用层的数据通信协议,构建全球工业系统统一标准的数据通信协议在短期内是不可能实现的,长远来看也会面临很多困难,而传统的数据采集技术无法从根本上解决上述技术问题。

发明内容

[0005] 针对现有技术的以上缺陷或改进需求,本发明提供一种面向协议数据流的数据抽取方法,既满足当前对协议数据流的数据抽取要求,又保证了数据抽取的通用性、灵活性以及可扩展性。

[0006] 本发明提供一种面向协议数据流的数据抽取方法,包括以下步骤:

[0007] 步骤 1 建立针对工业通信协议的数据抽取规则,所述数据抽取规则包括参数解析规则和模式信息规则,分别保存于参数解析规则文件和模式信息规则文件,其中,所述参数解析规则文件用于描述数据帧的类型与结构,所述模式信息规则文件用于描述数据转换与处理规则;

[0008] 步骤 2 读取所述参数解析规则文件,将其中所描述的不同解析节点实例化成具体的对象,生成由实例对象构成的数据解析树;读取所述模式信息规则文件,生成数据模式映射表;

[0009] 步骤 3 从协议数据流中接收数据帧，并利用所述数据解析树从所述数据帧中提取出关于数据项内容的描述性信息，以获取抽取数据项所需要的解析参数；

[0010] 步骤 4 利用所述解析参数以及所述数据模式映射表，确定所述数据帧中数据域的模式信息，其中，所述模式信息包括所述数据域的结构、格式和类型；

[0011] 步骤 5 根据所述解析参数以及所述模式信息，利用有限状态机实现对所述数据帧中数据项的抽取，并将抽取的数据项转换成结构化的数据保存。

[0012] 总体而言，通过本发明所构思的以上技术方案与现有技术相比，具有以下有益效果：

[0013] 本方法通过对现有工业中数据帧的共同特点，定义了数据帧的公共模型及其描述方法与参数，并允许相关领域的设计开发人员基于该模型对新的数据协议进行配置建模，而无需重新设计开发协议软件，最终实现数据帧的解析与数据抽取的目的。面向协议数据流的数据抽取方法可以对各种类型的协议数据帧的数据实现有效而准确地抽取，更能保证数据抽取的扩展性、通用性以及灵活性，即在协议改变的情况下该方法也能适用，显著提高了在工业领域中对二进制的数据帧数据抽取的普适性

附图说明

[0014] 图 1 为本发明实施例数据抽取的过程示意图；

[0015] 图 2 为本发明实施例树形解析节点的示意图；

[0016] 图 3 为本发明实施例数据解析树和数据模式映射表生成的过程示意图；

[0017] 图 4 为本发明实施例数据模式的结构示意图；

[0018] 图 5 为本发明实施例基于树形结构的参数解析的过程示意图；

[0019] 图 6 为本发明实施例基于有限状态机的数据项提取和结构化生成的过程示意图。

具体实施方式

[0020] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

[0021] 本发明包括以下三个部分：提取描述性信息，以获取抽取数据项所需要的解析参数，为数据域中数据的抽取做准备；利用上阶段所获取的解析参数，确定数据帧中数据域的模式信息，数据域的模式信息主要包括数据域的结构、格式和类型；根据解析参数和数据域的模式信息，实现对数据项的抽取，并转换成结构化的数据保存。

[0022] 图 1 所示为本发明实施例中数据抽取的过程示意图，具体包括以下步骤：

[0023] 步骤 1 建立针对实际应用工业通信协议的数据抽取规则，包括参数解析规则和模式信息规则，该两类规则分别保存于参数解析规则文件和模式信息规则文件。其中，参数解析规则文件主要用于描述数据协议的数据帧的类型与结构，例如某一种类型的数据帧包含什么数据内容，偏移地址、长度等信息；模式信息规则文件主要用于描述数据转换与处理规则，包括原子参数处理规则和组合参数处理规则，并包含了转换与处理函数名及相关参数。

[0024] 步骤 2 读取参数解析规则文件，将该参数解析规则文件中所描述的不同节点实例

化成具体的对象，在内存中构建由实例对象构成的数据解析树；读取模式信息规则文件，在内存中构建数据模式映射表。

[0025] 步骤3从协议数据流中接收数据帧，并利用数据解析树从数据帧中提取出关于数据项内容的描述性信息，获取抽取数据项所需要的解析参数，为数据域中数据内容的抽取做准备。其中，协议数据流是指按照步骤1中提及的工业通信协议建立的数据通信通道，数据帧为数据通信的基本单元。

[0026] 步骤4利用步骤3所获取的解析参数以及步骤2生成的数据模式映射表，确定该数据帧中用于保存待抽取数据的数据域的模式信息，其中，数据域的模式信息主要包括数据域的结构、格式和类型。

[0027] 步骤5根据步骤3获取的解析参数以及步骤4获取的数据域的模式信息，利用有限状态机实现对数据帧中的数据项的抽取，并将抽取的数据项转换成结构化的数据保存。

[0028] 本发明数据抽取方法的关键在于数据抽取规则，其中数据抽取规则的参数解析规则和模式信息规则由相应配置文件信息生成。在本发明实施例中，通过XML格式对参数解析规则文件和模式信息规则文件进行内容管理。

[0029] 下面将分别介绍参数解析规则文件和模式信息规则文件的格式。

[0030] 在本发明实施例中，参数解析规则文件将通过如下单个解析节点的配置示例以说明：

[0031]

```
<ParaConfig> //配置
    <Node Information> //解析节点信息
        <Node property> //解析节点属性
            <Node Type></Node Type> //解析节点的参数解析类型
            <Node Key></Node Key> //解析节点关键字
            <Node Name></Node Name> //解析节点名称
        </Node property>
        <Extraction Parameter> //抽取参数
            <Address Distribution></Address Distribution>
//位置分布类型
        <Begin Address ></Begin Address > //起始地址
        <Length></Length> //长度
        <Data></Data> //逻辑与操作数
        <Right offset></Right offset> //右偏移位数
        <Data Type></Data Type> //目标数据格式
    </Extraction Parameter>
    <Node selection> //解析节点选择
        <Child Node> //解析节点
            <Node Key></Node Key> //解析节点关键字
            <Node Name></Node Name> //解析节点名称
            <Function></Function> //解析函数
            <Extraction Parameter> //抽取参数
            .....
        </Extraction Parameter>
    </Child Node>
```

[0032]

.....

</Node selection>

</Node Information>

</ParaConfig>

[0033] 在本发明实施例中，模式信息规则文件通过如下单个数据模式的配置示例以说明：

[0034]

<Data Model> //数据模式

<Identification></Identification> //模式标识

<Attributes List> //属性链表

<Attributes Node> //元组属性信息

<Attributes Mark></Attributes Mark> //属性标识

<Function></Function> //操作函数

<ParameterSet> //参数集

<Begin Address></Begin Address> //起始地址

<Length></Length> //长度

</ParameterSet>

</Attributes Node>

.....

</Attributes List>

</Data Model>

[0035] 本发明中的参数解析方法的生成关键是参数解析和解析节点队列采用基于树形的数据结构，以下将详细说明参数解析和解析节点队列的相关参数。

[0036] 在本发明实施例中，根据参数解析规则的结构形式将参数解析规则分为两种主要类型：原子参数解析结构以及组合参数解析结构。

[0037] 其中，原子参数解析结构包含六个主要属性，采用六元组来描述： $In = \langle Ad, Be, L, D, Rn, DT \rangle$ ，其中 In 表示原子参数解析结构； Ad 表示位置分布类型； Be 表示语义单元起始地址； L 表示原子语义单元的长度； D 表示与源数据实现逻辑与的操作数； Rn 表示需要右偏移位数； DT 表示目标数据格式。

[0038] 组合参数解析结构，包含两个主要属性，采用二元组来描述： $InS = \langle G, Lin \rangle$ ，其中

InS 表示组合参数解析结构 ;G 表示组合规则, 即通过 G 将多个原子信息单元组合成有实际意义的信息单元 ;Lin 表示参数解析集, 包含了多个配置属性如上述的原子参数解析结构 In。

[0039] 在本发明实施例中, 解析节点选择方式主要分为两种形式 : 基于特征字选择方式和基于令牌选择方式。

[0040] 其中, 基于特征字选择方式是一种基于特征字查询的解析节点选择方式, 包含一个哈希映射表用于特征字的快速查询。该哈希映射表包含了所有解析节点特征字的哈希表, 可用二元组 (key, value) 表示, 其中 key 对应解析节点的特征字, value 为 key 对应的解析节点的名称。

[0041] 基于令牌选择方式是一种利用解析参数值作为判断依据查找解析节点的方式, 包含一个模式信息链表, 链表中的每个解析节点由三部分组成 : 逻辑算子 F, 通过该逻辑算子 F 以及抽取参数集 P 中的对应参数来判断是否选择当前的解析节点 ; 抽取参数集 P, 为与当前逻辑算子 F 相对应的参数 ; 节点名称 N, 表示当前解析节点对应的解析节点名称。

[0042] 图 2 所示为本发明实施例步骤 2 中数据解析树和数据模式映射表生成的过程示意图, 具体包括以下子步骤 :

[0043] 在步骤 201 中, 获取节点的配置信息, 读取参数解析规则文件, 依次读取文件中的配置信息并转换为 XML 格式对应的参数解析规则 DOM 对象 ;

[0044] 在步骤 202 中, 根据参数解析规则 DOM 对象中节点的 <Node Type> 判断当前节点的参数解析类型, 如果只有一个原子解析参数, 则执行步骤 203, 否则表示存在多个原子解析参数, 执行步骤 204 ;

[0045] 在步骤 203 中, 读取参数解析规则 DOM 对象中的 <Address Distribution>、<Begin Address>、<Length>、<Data>、<Right offset> 和 <Data Type> 等参数, 并相应赋予位置分布类型 Ad、语义单元的起始地址 Be、原子语义单元的长度 L、与源数据实现逻辑与的操作数 D、右偏移位数 Rn、目标数据格式 DT, 完成原子参数解析结构 In 的初始化, 然后执行步骤 205 ;

[0046] 在步骤 204 中, 从参数解析规则 DOM 对象中, 按照步骤 203 的方式读取多个原子参数解析结构, 并构建组合参数解析结构, 将读取的多个原子参数解析结构赋予参数解析集 Lin, 并进一步读取组合规则 G 以及参数解析集 ;

[0047] 在步骤 205 中, 根据参数解析规则 DOM 对象中 <Extraction Parameter> 标签下各子标签内容是否为空来判断当前节点的解析节点选择方式, 如果不为空则当前节点采用基于特征字选择方式, 执行步骤 206, 否则执行步骤 207, 即当前节点采用基于令牌选择方式 ;

[0048] 在步骤 206 中, 创建基于哈希表的解析节点映射表, 将解析节点名称 N 作为 Key 值, 用在内存中建立的对应解析节点对象 Op 的引用作为 Value 值, 构建 (Key, Value) 对插入到解析节点哈希映射表中, 继续执行步骤 208 ;

[0049] 在步骤 207 中, 依次读取解析节点名称 N、逻辑算子 F 以及抽取参数集 P, 并将其按参数解析规则文件中的顺序插入到模式信息链表中 ;

[0050] 在步骤 208 中, 判断参数解析规则文件是否读取完成, 如果是则执行步骤 209, 否则执行步骤 201 ;

[0051] 在步骤 209 中, 读取模式信息规则文件, 依次读取该文件中各个数据项的模式信息 ;

[0052] 在步骤 210 中, 读取模式信息中的各个元组信息, 并采用链式结构将其存储于内存中, 具体结构如图 3 所示。图 3 所示为本发明实施例的数据模式的结构示意图, 模式信息在内存中的存储结构为一链表——模式信息链表, 链表的每个节点均包含属性标识、操作函数及参数集, 该参数集进一步包含数据帧中对应数据项的偏移地址和长度;

[0053] 在步骤 211 中, 以模式标识 Mn 作为 Key 值, 对应的模式信息链表入口地址作为 Value 值建立基于哈希表的数据模式映射表;

[0054] 在步骤 212 中, 判断模式信息规则文件是否读取完毕, 如果是则本步骤结束, 执行步骤 3, 否则继续执行步骤 209。

[0055] 图 4 所示为本发明实施例树形解析节点的示意图, 采用多叉树结构, 每个父节点包含有多个子节点。

[0056] 图 5 所示为本发明实施例步骤 3 中基于树形结构的参数解析的过程示意图, 具体包括以下子步骤:

[0057] 在步骤 301 中, 接收待解析的数据帧;

[0058] 在步骤 302 中, 根据解析树根节点的名称到解析节点对象映射表中获取根解析节点对象 Op;

[0059] 在步骤 303 中, 执行当前解析节点的解析函数, 判断当前参数解析类型, 如果是原子参数类型, 则执行步骤 304, 否则执行步骤 307;

[0060] 在步骤 304 中, 读取解析节点对象 Op 中配置的抽取参数 Ep 信息, 并根据起始地址 Be 和长度 L 定位到当前参数的数据帧区域, 获取数据项内容;

[0061] 在步骤 305 中, 用逻辑与操作数 D 与步骤 304 中获取的数据项内容做逻辑与操作, 将结果再按右偏移位数 Rn 向右偏移, 获取参数在数据区域的有效位;

[0062] 在步骤 306 中, 根据目标数据格式 DT 将从数据帧中获取的数据转换成目标数据, 然后执行步骤 312;

[0063] 在步骤 307 中, 遍历模式信息链表, 依次读取抽取参数 Ep 的信息;

[0064] 在步骤 308 中, 根据读取的抽取参数 Ep 中的起始地址 Be 和长度 L 定位到当前参数的数据帧区域并获取数据项内容, 用逻辑与操作数 D 与获取的数据项内容做逻辑与操作, 将结果再按右偏移位数 Rn 向右偏移, 获取参数在数据区域的有效位, 根据目标数据格式 DT 将从数据帧中获取的数据转换成目标数据;

[0065] 在步骤 309 中, 将步骤 308 抽取的参数存储到一个临时的抽取参数链表;

[0066] 在步骤 310 中, 判断是否到达模式信息链表尾, 是则执行 311, 否则执行 307;

[0067] 在步骤 311 中, 根据组合规则 G 以及已获取的相关参数, 得到组合参数数据;

[0068] 在步骤 312 中, 根据参数名称 Ps(来自模式信息规则文件中的<ParameterSet>标签下的参数), 将获取的参数值输出到共享参数列表中;

[0069] 在步骤 313 中, 判断当前节点的选择方式, 如果是基于特征字的选择方式, 则执行步骤 314, 否则执行步骤 315, 即为基于令牌的选择方式;

[0070] 在步骤 314 中, 根据步骤 312 获取的参数值, 从解析节点映射表中获取对应解析节点的名称, 然后执行步骤 316;

[0071] 在步骤 315 中, 根据获取的参数值、抽取参数以及逻辑算子, 依次遍历模式信息链表, 直到得到节点名称;

[0072] 在步骤 316 中,通过节点名称,获取下一节点的解析节点对象 Op ;

[0073] 在步骤 317 中,判断当前是否到达解析树的叶子节点,是则参数解析过程结束,否则执行步骤 303。

[0074] 本发明实施例中提供的数据项提取和结构化生成方法核心是利用有限状态机来实现对数据项的数据提取,以下将对本发明中的有限状态机模型做详细说明。

[0075] 有限状态机的包括三个主要方面 :转移状态、输入事件以及转移函数。

[0076] 有限状态机转移状态 :状态机的状态集 $Q = \{B, M, V, Q, T, E\}$, 包括六个状态分别为 B、M、V、Q、T 和 E, 其中 B 表示起始状态, 表示数据域数据抽取的开始状态或者是数据域中某个数据记录抽取的开始状态 ;E 表示结束状态, 表示当前数据域的数据抽取过程全部完成 ;M、V、Q、T 状态分别表示提取完数据标识、数据元素值、数据质量以及数据时间后所处的状态。

[0077] 有限状态机的输入事件 :状态机有穷的输入符号集合 $\Sigma = \{b, m, v, q, t, e\}$, 包括六个输入事件 b、m、v、q、t 和 e, 其中 b 事件表示抽取完记录的最后一个元组数据, 但当前数据域抽取未结束, 数据域中数据记录的个数表示为 N, 每次抽取完一条数据记录后 N 递减 1, 根据 N 的值来判断数据域抽取是否结束, 当 N 大于 0 时则表示抽取未完。反之, 当 N 等于 0 表示抽取结束, 事件 b 可以表述为 $b = \langle le, N1 \rangle$, le 表示最后元组事件, N1 表示 N 取值大于 0 ;e 事件表示抽取完成后记录的最后一个元组数据, 并且当前数据域抽取已经全部完成, $e = \langle le, N0 \rangle$, le 如 b 中表述一样, N0 表示 N 等于 0 事件 ;事件 m、v、q、t 分别表示下一阶段的要实际抽取的数据为数据标识、数据元素值、数据质量以及数据时间。有限状态机通过循环遍历数据模式映射表, 分别生成数据模式标识所对应的事件, 当遍历到链表尾时产生事件 1。事件 e 表示收取结束。

[0078] 有限状态机的转移函数 :有限状态机的状态转移函数主要是实现数据提取以及缺失属性的数据补齐。数据提取是通过当前事件所对应的二元组数据提取参数 $Me = \langle OF, OP \rangle$, 其中 OF 表示操作函数, OP 表示函数参数集, 从数据域中提取当前数据记录的元组属性, 该元组属性来自模式信息规则文件的 <Attributes Node> 标签包含的所有参数, 对于函数参数集 OP 中的起始地址是当前元组属性相对当前属性所在数据项的偏移地址, 在数据抽取过程要不断更新数据记录的起始地址, 定义 $L = L0 + n \times Ld$, 其中 L0 表示数据域的起始偏移, n 表示当前已抽取的数据记录数, Ld 表示数据记录的长度, 对于特定类型的数据模式, 其数据记录的长度是相同的。例如对于状态转移函数 $S(B, v)$ 所描述的过程即为当前状态为 B 接收到事件 v, 状态要转移到 V 状态, 在元组属性完整的情况下, B 状态首先应该收到的应该是 m 事件, 即数据标识事件, 所以这里就存在着数据标识属性的缺失, 需要补齐该数据记录中的数据标识。同样对于 $S(V, b)$ 则缺失了数据质量和数据时标, 需要对应补齐。缺失的属性取值有两种形式 :数据预处理阶段的解析参数和默认参数。即首先在解析参数表中查找是否有对应的属性值, 如果没有找到则通过读取默认的参数, 例如对于时标的默认参数即为系统当前时间。对于数据标识属性缺失, 其属性值 $V = f(Vb, np)$ 来表示, Vb 为数据起始标识, np 表示当前已经处理数据记录, f 表示对 Vb 和 np 的数据运算, 运算函数 f 由具体的数据协议的数据项定义, 具体以回调函数的形式配置到模式信息规则文件中。

[0079] 图 6 所示为本发明实施例步骤 5 中基于有限状态机的数据项提取和结构化生成的过程示意图, 具体包括以下步骤 :

[0080] 在步骤 501 中,根据上述步骤 3 生成的参数,获取数据项的抽取参数,例如起始偏移地址 Offset、数据项个数 N 以及数据时标 Time 等信息;

[0081] 在步骤 502 中,从起始偏移地址 Offset 处开始读取数据帧,定位数据域的初始地址,并将数据模式偏移到首地址处;

[0082] 在步骤 503 中,根据数据模式的偏移地址获取到当前的输入事件以及元组提取参数,并执行步骤 504;

[0083] 在步骤 504 中,判断当前所处状态以及输入事件,判断条件用 S_{Se} 表示,如下表 1 中所示,其中下标 S 表示当前所处状态, e 表示当前输入事件,例如表 1 中 S_{Bv} , 则表示当前状态处在初始 B 状态,接收到 v 事件。如果当前为 S_{Bm} , 则执行步骤 505, 如果当前为 S_{Bv} , 则执行步骤 506, 如果当前为 S_{Mv} , 则执行步骤 507, 如果当前为 S_{Vq} , 则执行步骤 508, 如果当前为 S_{Vt} , 则执行步骤 509, 如果当前为 S_{Qt} , 则执行步骤 510, 如果当前为 S_{Ve} , 则执行步骤 511, 如果当前为 S_{Qe} , 则执行步骤 512, 如果当前为 S_{Te} , 则执行步骤 513, 如果当前 S_{Tb} , 则执行步骤 514, 如果当前 S_{Qb} , 则执行步骤 515, 如果当前为 S_{vb} , 则执行步骤 516;

[0084]

事件 状态	b	m	v	q	t	e
B	-	S_{Bm}	S_{Bv}	-	-	-
M	-	-	S_{Mv}	-	-	-
V	S_{vb}	-	-	S_{Vq}	S_{Vt}	-
Q	S_{Qb}	-	-	-	-	S_{Qe}
T	S_{Tb}	-	-	-	-	S_{Te}
E	-	-	-	-	-	-

[0085] 表 1

[0086] 在步骤 505 中,通过数据标识的操作函数 OF 和函数参数集 OP,从指定偏移处提取出数据项中的数据标识,数据项循环个数减 1,有限状态机状态转移到状态 M,并执行步骤 517;

[0087] 在步骤 506 中,根据解析参数中的数据项基址 Ba 和当前已抽取数据项个数 ND,得到数据记录地址,补齐当前数据记录的数据地址,通过数据值 v 的提取操作函数 OF 和函数参数集 OP,从指定偏移处提取出数据项中的数据值 v,数据项循环个数减 1,有限状态机状态转移到状态 V,并执行步骤 517;

[0088] 在步骤 507 中,通过数据值 v 的提取操作函数 OF 和函数参数集 OP,从指定偏移处提取出数据项中的数据值 v,有限状态机状态转移到状态 V,并执行步骤 517;

[0089] 在步骤 508 中,通过数据质量 q 的提取操作函数 OF 和函数参数集 OP,从指定偏移处提取出数据项中的数据质量 q,有限状态机状态转移到状态 Q,并执行步骤 517;

[0090] 在步骤 509 中,从解析参数中获取到当前数据项的数据质量 q,补齐当前数据记录

的数据质量 q, 通过数据时间 t 的提取操作函数 OF 和函数参数集 OP, 从指定偏移处提取出数据项中的数据时间 t, 有限状态机状态转移到状态 T, 并执行步骤 517;

[0091] 在步骤 510 中, 通过数据时间 t 的提取操作函数 OF 和函数参数集 OP, 从指定偏移处提取出数据项中的数据时间 t, 有限状态机状态转移到状态 T, 并执行步骤 517;

[0092] 在步骤 511 中, 从解析参数中获取到当前数据项的数据质量 q, 补齐当前数据记录的数据质量 q, 从解析参数获取到该数据帧的数据时间 t, 补齐当前数据记录的数据时间, 有限状态机状态转移到终止状态 E, 结束流程;

[0093] 在步骤 512 中, 从解析参数获取到该数据帧的数据时间 t, 补齐当前数据记录的数据时间, 有限状态机状态转移到终止状态 E, 结束流程;

[0094] 在步骤 513 中, 有限状态机状态转移到终止状态 E, 结束流程;

[0095] 在步骤 514 中, 有限状态机状态转移到开始状态 B, 并执行步骤 517;

[0096] 在步骤 515 中, 从解析参数获取到该数据帧的数据时间 t, 补齐当前数据记录的数据时间, 有限状态机状态转移到开始状态 B, 并执行步骤 517;

[0097] 在步骤 516 中, 从解析参数中获取到当前数据项的数据质量 q, 补齐当前数据记录的数据质量, 从解析参数获取到该数据帧的数据时间 t, 补齐当前数据记录的数据时间, 有限状态机状态转移到开始状态 B, 并执行步骤 517;

[0098] 在步骤 517 中, 判断当前是否为数据项的最后一个元组, 如果是则执行步骤 518, 否则执行步骤 519;

[0099] 在步骤 518 中, 判断当前数据项个数是否为 0, 如果是则执行步骤 520, 否则执行步骤 521;

[0100] 在步骤 519 中, 数据模式偏移到下一元组属性地址处, 并执行步骤 503;

[0101] 在步骤 520 中, 生成结束事件 e, 执行步骤 504;

[0102] 在步骤 521 中, 数据模式偏移到首地址处, 生成开始事件 b, 并执行步骤 504。

[0103] 本领域的技术人员容易理解, 以上所述仅为本发明的较佳实施例而已, 并不用以限制本发明, 凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等, 均应包含在本发明的保护范围之内。

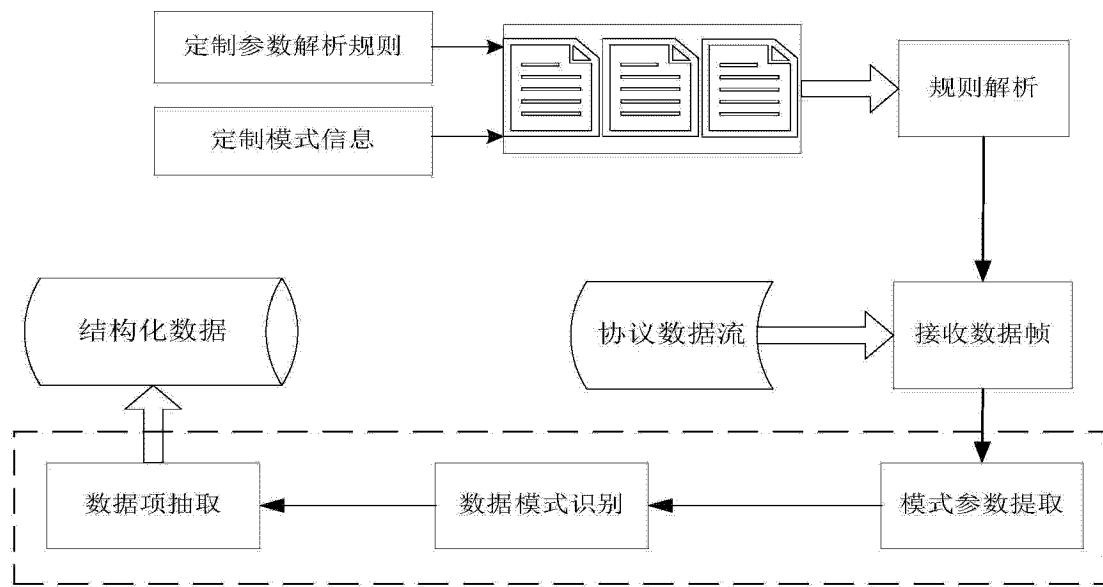


图 1

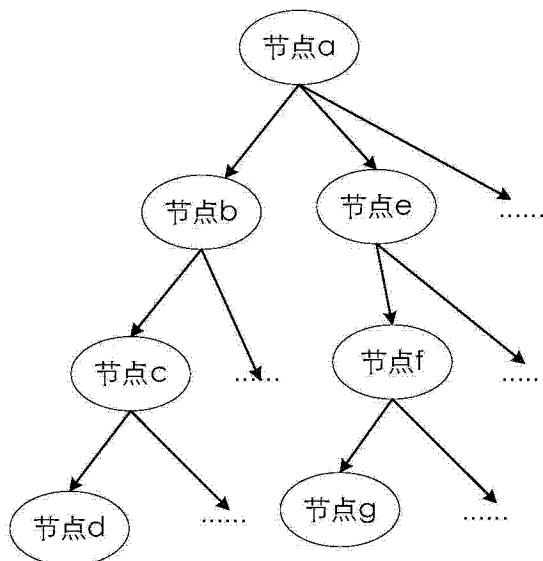


图 2

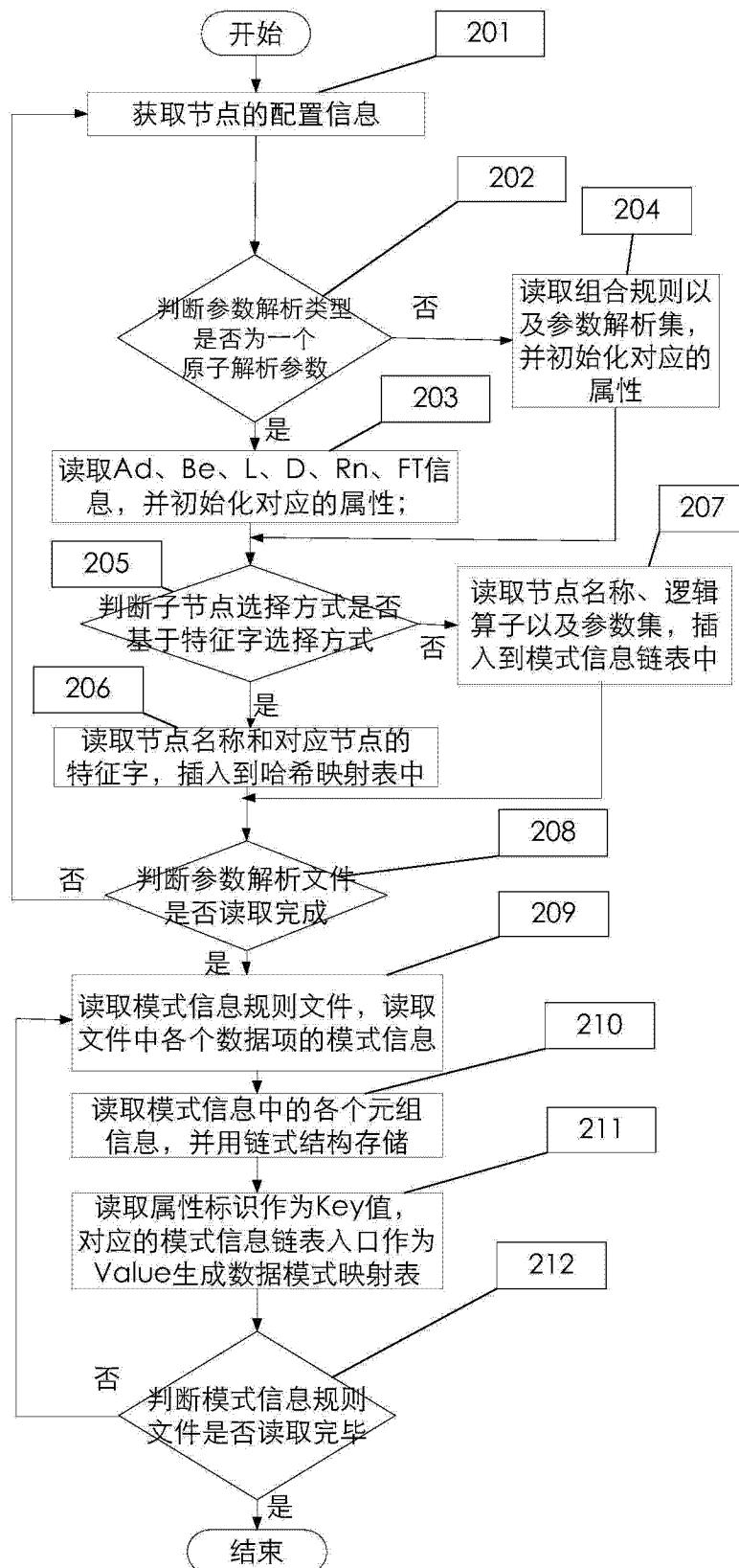


图 3



图 4

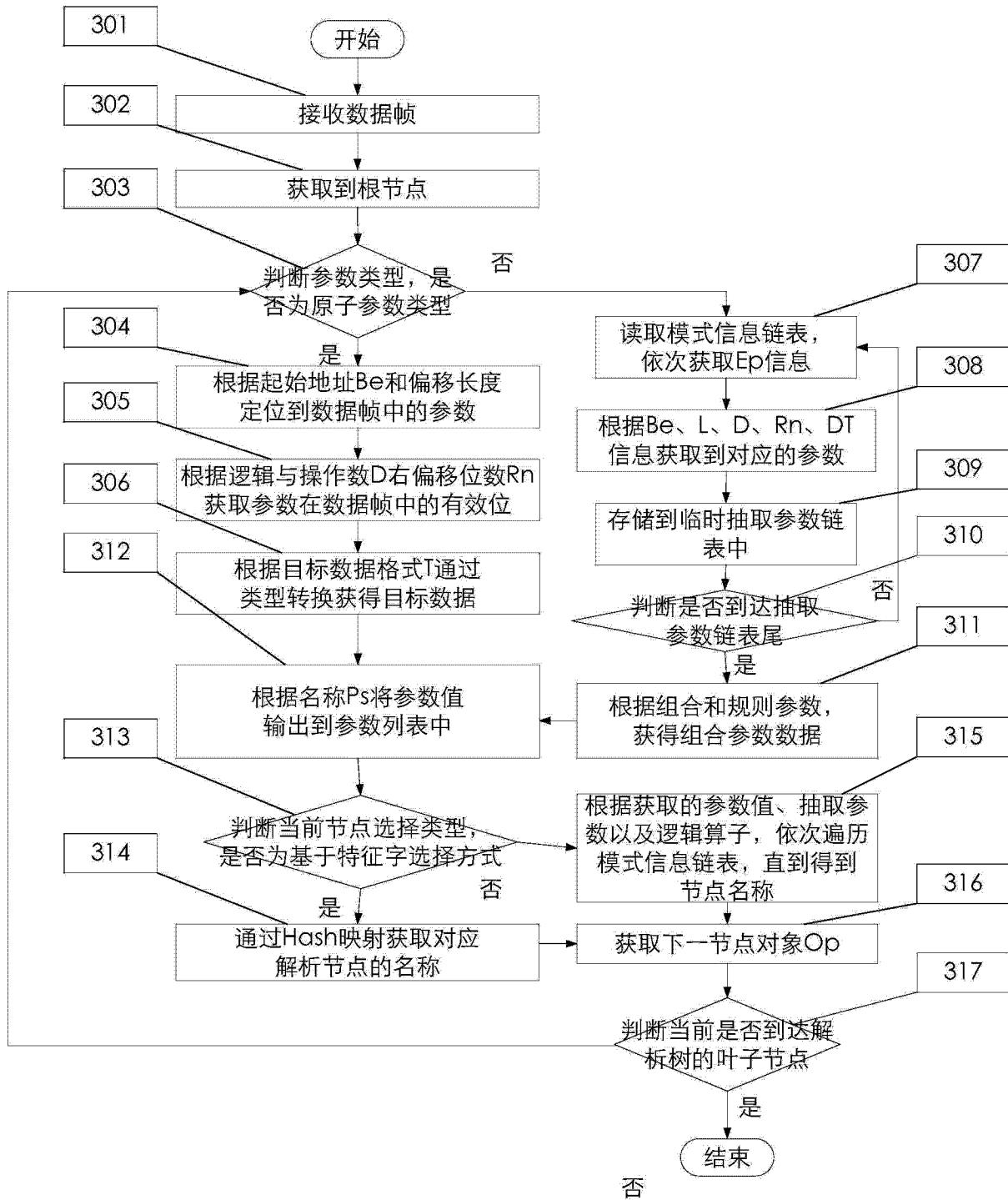


图 5

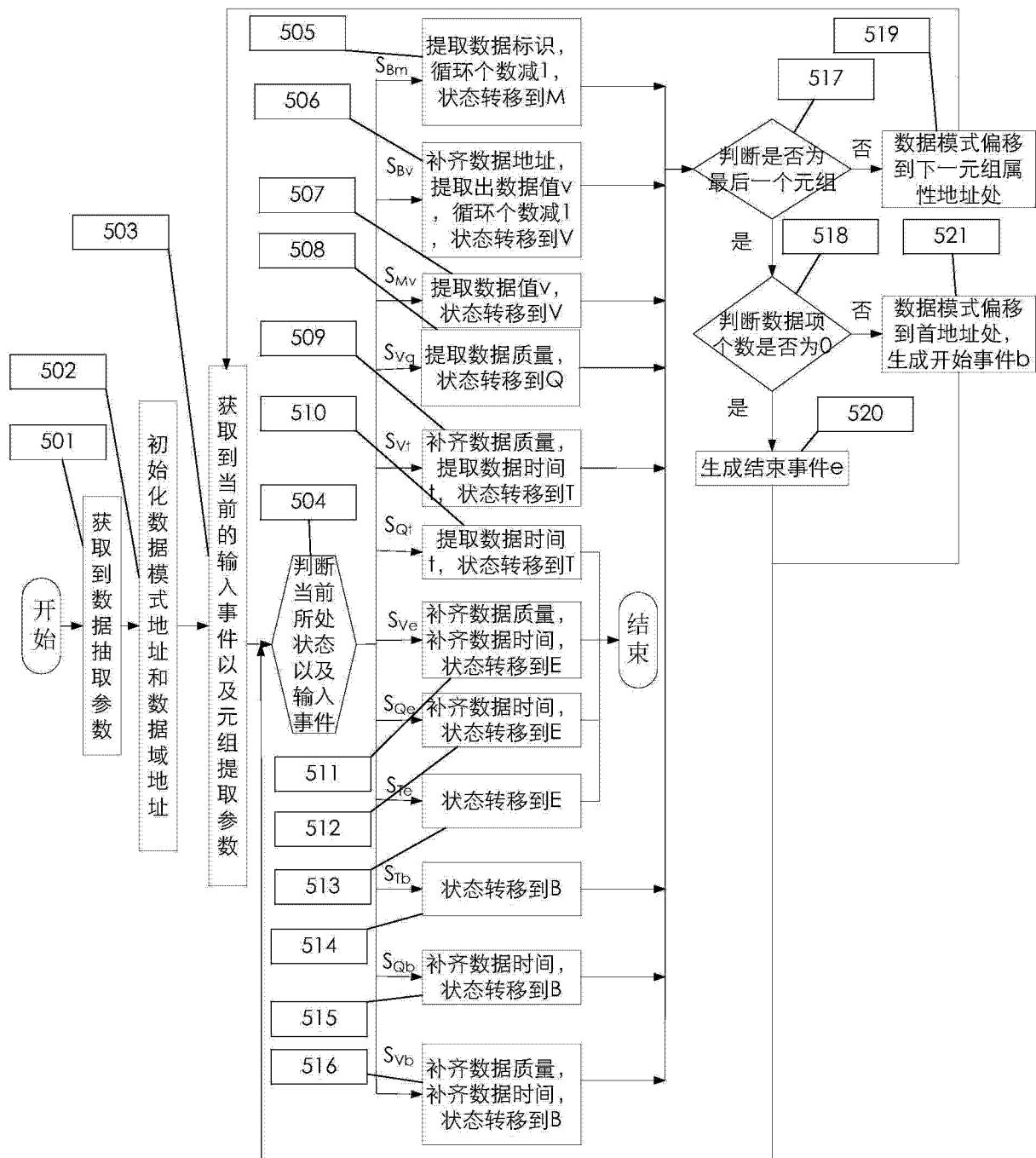


图 6