(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117971819 B
(45)授权公告日 2024.05.31

(21)申请号 202410371733.X

(22)申请日 2024.03.29

(65)同一申请的已公布的文献号
申请公布号 CN 117971819 A

(43)申请公布日 2024.05.03

(73)专利权人 南京金鼎嘉崎信息科技有限公司
地址 210008 江苏省南京市雨花台区软件大道11号花神大厦207、208、209室

(72)发明人 武春庆

(74)专利代理机构 南京佰腾智信知识产权代理事务所(普通合伙) 32509
专利代理师 郭林

(51)Int.Cl.
G06F 16/215(2019.01)
G06F 16/22(2019.01)
G06F 16/2455(2019.01)
G06F 18/22(2023.01)

(56)对比文件
CN 109146663 A,2019.01.04
CN 110020303 A,2019.07.16
CN 111538733 A,2020.08.14
CN 113901768 A,2022.01.07
CN 114841806 A,2022.08.02
CN 115526722 A,2022.12.27
US 10911583 B1,2021.02.02
周赟.基于JPEG2000的自适应算术编解码器的研究与实现.《中国优秀硕士学位论文全文数据库 信息科技辑》.2008,I135-107.
Hong Y 等.Use of satellite remote sensing data in the mapping of global landslide susceptibility.《Natural hazards》.2007,245-256.

审查员 杨春颖

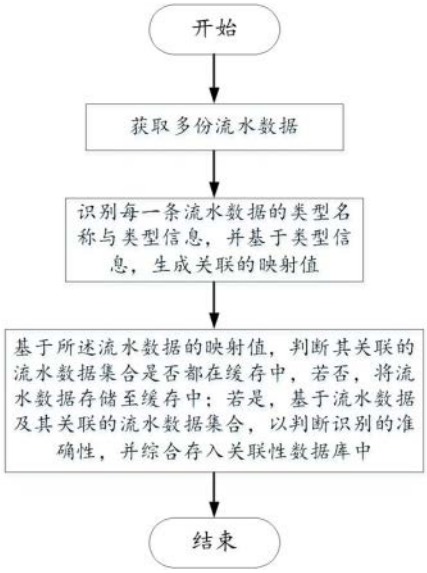
权利要求书2页 说明书6页 附图1页

(54)发明名称

自动汇集流水数据的管理方法与系统

(57)摘要

一种自动汇集流水数据的管理方法,包括:获取多份流水数据;识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中,若否,将流水数据存储至缓存中;若是,基于流水数据及其关联的流水数据集合,以判断识别的准确性,并综合存入关联性数据库中。本发明采取了一种新的自动汇集流水数据的方法,不同于传统方式中只采用正则表达式制式的对流水数据进行识别,或者采用自然语言处理进行识别。本方法具有极高的识别准确率。



CN 117971819 B

1.一种自动汇集流水数据的管理方法,其特征在于,所述方法包括步骤1~步骤3;

步骤1,获取多份流水数据;

步骤2,识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;基于类型信息,生成关联的映射值具体为:基于类型信息中的交易时间,生成第一映射值,并基于类型信息中的交易对方信息或交易平台信息中的多个稀缺关键字,生成多个第二映射值;其中,稀缺关键字指的是一个汉字或一个单词;

选取稀缺关键字的过程具体包括步骤S101~步骤S103;

步骤S101,获取所述流水数据中类型信息中的交易对方信息或交易平台信息中的的每一个关键字;

步骤S102,根据哈夫曼编码,计算出每一个关键字对应的映射值;

步骤S103,从高到低对映射值进行排序,选取映射值最大的n个,作为多个第二映射值,其中,n为第二映射值的数量;

步骤3,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中,若否,将流水数据存储至缓存中;若是,基于流水数据及其关联的流水数据集合,以判断识别的准确性,并综合存入关联性数据库中;

步骤3中,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中具体包括步骤3.1~步骤3.3;

步骤3.1,基于第一映射值,获取第一后继映射值;

步骤3.2,基于所述流水数据的第一映射值,获取第一映射值与第一后继映射值中的所有流水数据作为待比较流水数据集;

步骤3.3,将所述流水数据的第二映射值依次与待比较流水数据集中每一个流水数据的第二映射值进行比较,若二者相似度大于等于预设的相似度阈值,则判定所述每一个流水数据为关联的流水数据。

2.根据权利要求1所述的一种自动汇集流水数据的管理方法,其特征在于,在获取多份流水数据前还包括对多份流水数据进行预处理,包括:数据清洗与标准化。

3.根据权利要求1所述的一种自动汇集流水数据的管理方法,其特征在于,步骤2中通过正则表达式的方式实现对流水数据的识别分割。

4.一种自动汇集流水数据的管理系统,应用于权利要求1~3任一所述的方法上,其特征在于,所述系统包括:数据获取模块、逻辑判断模块、第一数据存储模块与第二数据存储模块;

数据获取模块用于获取多份流水数据;

逻辑判断模块用于识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;以及基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中;以及基于流水数据及其关联的流水数据集合,以判断识别的准确性;

第一数据存储模块用于将流水数据存储至缓存中;

第二数据存储模块用于将综合存入关联性数据库中。

5.一种终端,包括处理器及存储介质;其特征在于:

所述存储介质用于存储指令;

所述处理器用于根据所述指令进行操作以执行根据权利要求1-3任一项所述方法的步

骤。

6.计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现权利要求1-3任一项所述方法的步骤。

自动汇集流水数据的管理方法与系统

技术领域

[0001] 本发明属于数据分析技术领域,更具体的,涉及一种自动汇集流水数据的管理方法与系统。

背景技术

[0002] 在现代经济活动中,流水数据,如银行交易记录、电子支付信息和通信话单,扮演着至关重要的角色。这些数据反映了个人和企业的财务状况、消费习惯和沟通模式。然而,由于这些信息通常以非结构化文本形式出现,使得数据的整合、分析和管理变得极其复杂和耗时。当前的背景下,企业和个人需要从各种源收集和整合流水数据,以支持决策制定、财务管理和客户服务等关键业务活动。

[0003] 目前,自动汇集流水数据的技术主要依赖于正则表达式、自然语言处理(NLP)、数据挖掘和机器学习等方法。通过实现高级的实体识别和模式匹配算法,系统能够识别文本中的关键信息,如日期、金额和参与方等。此外,文本分类和情感分析技术被用于进一步分析数据,从而为用户提供洞察力。这些技术的整合形成了能够自动处理和汇总流水数据的系统,为用户提供了一个更加清晰、有序的财务画面。

[0004] 尽管有显著进步,当前的自动汇集流水数据技术仍然存在一些不足之处。首先,制式的正则表达式对复杂和非标准化文本的处理仍然是一个挑战。其次,自然语言处理对于文本中的隐含意义和上下文信息的理解还不够深入,这限制了信息提取的准确性和系统的应用范围。

发明内容

[0005] 为解决现有技术中存在的不足,本发明的目的在于解决上述缺陷,进而提出一种自动汇集流水数据的管理方法与系统。

[0006] 本发明采用如下的技术方案。

[0007] 本发明第一方面公开了一种自动汇集流水数据的管理方法,包括步骤1~步骤3;

[0008] 步骤1,获取多份流水数据;

[0009] 步骤2,识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;

[0010] 步骤3,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中,若否,将流水数据存储至缓存中;若是,基于流水数据及其关联的流水数据集合,以判断识别的准确性,并综合存入关联性数据库中。

[0011] 进一步的,在获取多份流水数据前还包括对多份流水数据进行预处理,包括:数据清洗与标准化。

[0012] 进一步的,基于类型信息,生成关联的映射值具体为:基于类型信息中的交易对方信息或交易平台信息,生成关联的映射值。

[0013] 进一步的,步骤2中通过正则表达式的方式实现对流水数据的识别分割。

[0014] 进一步的,基于类型信息,生成关联的映射值具体为:基于类型信息中的交易时间,生成第一映射值,并基于类型信息中的交易对方信息或交易平台信息中的多个稀缺关键字,生成多个第二映射值;其中,稀缺关键字指的是一个汉字或一个单词。

[0015] 进一步的,步骤3中,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中具体包括步骤3.1~步骤3.3;

[0016] 步骤3.1,基于第一映射值,获取第一后继映射值;

[0017] 步骤3.2,基于所述流水数据的第一映射值,获取第一映射值与第一后继映射值中的所有流水数据作为待比较流水数据集;

[0018] 步骤3.3,将所述流水数据的第二映射值依次与待比较流水数据集中每一个流水数据的第二映射值进行比较,若二者相似度大于等于预设的相似度阈值,则判定所述每一个流水数据为关联的流水数据。

[0019] 进一步的,选取稀缺关键字的过程可以具体包括步骤S101~步骤S103;

[0020] 步骤S101,获取所述流水数据中类型信息中的交易对方信息或交易平台信息中的的每一个关键字;

[0021] 步骤S102,根据哈夫曼编码,计算出每一个关键字对应的映射值;

[0022] 步骤S103,从高到低对映射值进行排序,选取映射值最大的n个,作为多个第二映射值,其中,n为第二映射值的数量。

[0023] 本发明第二方面公开了一种自动汇集流水数据的管理系统,应用于第一方面所述的方法上,包括:数据获取模块、逻辑判断模块、第一数据存储模块与第二数据存储模块;

[0024] 数据获取模块用于获取多份流水数据;

[0025] 逻辑判断模块用于识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;以及基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中;以及基于流水数据及其关联的流水数据集合,以判断识别的准确性;

[0026] 第一数据存储模块用于将流水数据存储至缓存中;

[0027] 第二数据存储模块用于将综合存入关联性数据库中。

[0028] 本发明第三方面公开了一种终端,包括处理器及存储介质;其特征在于:

[0029] 所述存储介质用于存储指令;

[0030] 所述处理器用于根据所述指令进行操作以执行第一方面所述方法的步骤。

[0031] 本发明第四方面公开了一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现第一方面所述方法的步骤。

[0032] 本发明的有益效果在于,与现有技术相比,本发明具有以下优点:

[0033] 本发明采取了一种新的自动汇集流水数据的方法,不同于传统方式中只采用正则表达式制式的对流水数据进行识别,或者采用自然语言处理进行识别。本方法具有极高的识别准确率。

附图说明

[0034] 图1是本发明实施例的一种自动汇集流水数据的管理方法的流程图。

具体实施方式

[0035] 下面结合附图对本申请作进一步描述。以下实施例仅用于更加清楚地说明本发明的技术方案,而不能以此来限制本申请的保护范围。

[0036] 本发明公开了一种自动汇集流水数据的管理方法,如图1所示,可以包括步骤1~步骤3。

[0037] 步骤1,获取多份流水数据。

[0038] 其中,流水数据可以是银行账号、电子支付平台(例如:支付宝、微信等)、或其他通信服务商的交易数据。其中,流水数据为包含多种格式和结构的文本数据,通常,

[0039] 在一些实施例中,在获取多份流水数据前还包括对多份流水数据进行预处理,包括:数据清洗与标准化。其中,数据清洗用于去除无关信息,减少流水数据的数量。标准化用于将所有的流水数据统一格式化,同时识别出其中的关键信息,例如:交易日期、金额、交易方、交易描述等等,以便于进行后期的存储。

[0040] 步骤2,识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值。

[0041] 其中,实体信息可以包括交易时间、交易金额、交易类型、交易本方信息、交易对方信息、交易平台信息、订单号等。

[0042] 在本发明的第一实施例中,可以基于类型信息中的交易对方信息或交易平台信息,生成关联的映射值。假设一条流水数据X可以包含如下信息:交易本方信息:“x0”,交易对方信息:“x1”,则其关联的映射值仅为“x1”所对应的映射值;假设一条流水数据Y可以包含如下信息:交易本方信息:“y0”,交易对方信息:“y1”,交易平台信息:“y2”,则其关联的映射值为“y1”与“y2”所分别对应的2个映射值。

[0043] 在第一实施例中,所述映射值的生成方式可以通过hash函数的方式生成,其具体公式不再赘述。

[0044] 步骤3,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中,若否,将流水数据存储至缓存中;若是,基于流水数据及其关联的流水数据集合,以判断识别的准确性,并综合存入关联性数据库中。

[0045] 在一些实施例中,所述缓存可以选择redis数据库。

[0046] 可以理解的是,假设所述流水数据为流水数据X,则其关联的流水数据集合中流水数据的数量至少为1;假设所述流水数据为流水数据Y,则其关联的流水数据集合中流水数据的数量至少为2。步骤3中,综合存入关联性数据库中指的是不仅将所述流水数据存入关联性数据库中,还将其关联的流水数据集合从缓存中转移至关联性数据库中。

[0047] 在步骤3中,以流水数据Y为例,其对应的2个映射值分别关联的至少2个流水数据S、T与流水数据Y应当为同一个流水数据,仅仅是交易双方信息与交易平台信息发生了改变。

[0048] 在现代数据处理实践中,自动化地识别和汇集流水数据是一项重要且挑战性的任务。制式的机械识别每一条流水数据的实体信息,例如通过预设的正则表达式快速识别流水数据中的实体信息是一种常见方法,但这种方法存在其固有局限性。特别是,流水数据中包含的实体类型的顺序以及类型可能不尽相同,影响数据处理的一致性和准确性。此外,数据获取过程中的问题,如页码转换错误,可能导致流水记录被错误地合并或分割,或者与非

相关信息混合,进一步降低了自动化处理的准确度。即使自动汇集流水数据的成功率高达99%,事后的查缺补漏工作仍可能导致高昂的成本,从而抵消自动化带来的效率优势。

[0049] 在另一些实施例中,引入基于自然语言处理(NLP)的技术,可以提高对文本中实体的识别率,然而,其识别准确率甚至不如正则表达式高,同时自然语言处理也无法解决数据界定问题。也就是说,现有技术难以精确界定哪些内容确切属于特定的流水记录,或者根本不属于任何流水数据。

[0050] 在本发明的实施例中,由于步骤3通过交叉验证(即基于流水数据,及其关联的流水数据集合)的方式以判断识别的准确性,采用交叉验证的方式,能够轻易的解决识别准确率以及数据界定的问题。因此,步骤2中完全可以通过正则表达式的方式实现快速的对流水数据进行识别分割。然而,上述步骤中依然存在缺陷,首先最为重要的是,本发明映射值的目的本身就是为了能够准确的分类识别流水数据,而映射值本身又由流水数据中的类型信息(例如:交易对方信息或交易平台信息)唯一确定。因此,一旦通过正则表达式分类识别有误,则映射值就错了,这本质上就形成了一个先有鸡还是先有蛋的问题。其次,由于流水数据通常是分批(通常其交易本方信息是一样的)进行处理,不难理解的是,第一个批次的流水数据必然要全部存入缓存中,而通常缓存即为内存,空间上可能无法适应庞大的数据量。

[0051] 基于此,在本发明的第二实施例中,基于类型信息,生成关联的映射值具体为:基于类型信息中的交易时间,生成第一映射值,并基于类型信息中的交易对方信息或交易平台信息中的多个稀缺关键字,生成多个第二映射值;其中,稀缺关键字(以及下文中的关键字)指的是一个汉字或一个单词。

[0052] 采用交易时间来生成关联的映射值的优势在于,交易时间本身的格式是确定的,且通常流水数据的顺序按照交易时间进行排序,可以关联上下两条流水数据进行交叉验证从而获得准确的交易时间。

[0053] 更为具体的,考虑到所述流水数据与其关联的流水数据可能存在时间差Δt,因此,第一映射值应当涵盖的是整个映射区间(即下文中的[v, $v+d$])的流水数据。因此,在第二实施例中,所述第一映射值v可以为交易时间相关联的时间戳,如下式所示:

[0054] $$v = t\%d$$

[0055] 其中,$\%$为余运算符号,t为交易时间的时间戳,d为大于Δt的整数。

[0056] 可理解的,第二实施例中本质上利用了桶排序的思想,即:先根据交易时间生成第一映射值(本质上是一个映射区间,每一个映射区间看成一个木桶),大致确定所述流水数据在哪一个木桶中;然后再利用稀缺关键字,例如通过散列表来确定流水数据的具体位置。

[0057] 因此,步骤3中,基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中具体包括步骤3.1~步骤3.3。

[0058] 步骤3.1,基于第一映射值,获取第一后继映射值。

[0059] 第一后继映射值指的是第一映射值的下一个映射值。可理解的,第一后继映射值$v' = v + \mathrm{d}$。

[0060] 步骤3.2,基于所述流水数据的第一映射值,获取第一映射值与第一后继映射值中的所有流水数据作为待比较流水数据集。

[0061] 步骤3.3,将所述流水数据的第二映射值依次与待比较流水数据集中每一个流水数据的第二映射值进行比较,若二者相似度大于等于预设的相似度阈值,则判定所述每一个流水数据为关联的流水数据。

[0062] 可理解的,在缓存中,关联的流水数据的多个第二映射值应当也是从高到低进行排序,构成向量。通常,第二映射值的总数量n至少应当大于等于5,所述预设的相似度阈值可以设定为80%。所述二者相似度指的是两个流水数据的第二映射值中重复数据的个数与总数量的比值。

[0063] 需要说明的是,在步骤3.3中,若二者相似度大于等于预设的相似度阈值,通常还需要进一步比较两条流水数据信息,从而进一步判定所述每一个流水数据是否为关联的流水数据,然而这种判定方法无外乎是继续扩张第二映射值的数量的长度进行比较,以防止先前的n太小,导致判断不精准,相应的,当第二映射值的数量扩张时,相似度阈值也应当提高。具体过程不再赘述。

[0064] 为了防止稀缺关键字选取的是流水数据中公共的关键字,在本发明的实施例中,选取稀缺关键字的过程可以具体包括步骤S101~步骤S103。

[0065] 步骤S101,获取所述流水数据中类型信息中的交易对方信息或交易平台信息中的的每一个关键字。

[0066] 步骤S102,根据哈夫曼编码,计算出每一个关键字对应的映射值。

[0067] 步骤S103,从高到低对映射值进行排序,选取映射值最大的n个,作为多个第二映射值,其中,n为第二映射值的数量。

[0068] 可理解的,最大的n个第二映射值所对应的关键字即为稀缺关键字。

[0069] 需要说明的是,步骤S102中使用哈夫曼编码(Huffman Coding),并非是为了对数据进行压缩,而是因为哈夫曼编码的核心思想是根据每个关键字出现的频率或概率来分配不等长的位序列,即编码。在哈夫曼编码过程中,出现频率最高的关键字被赋予最短的映射编码,而出现频率低的关键字则被赋予较长的映射编码,因此,步骤S102中每一个关键字对应的映射值即为该关键字对应的哈夫曼编码中的映射编码。

[0070] 相应的,本发明还公开了一种自动汇集流水数据的管理系统,包括:数据获取模块、逻辑判断模块、第一数据存储模块与第二数据存储模块;

[0071] 数据获取模块用于获取多份流水数据;

[0072] 逻辑判断模块用于识别每一条流水数据的类型名称与类型信息,并基于类型信息,生成关联的映射值;以及基于所述流水数据的映射值,判断其关联的流水数据集合是否都在缓存中;以及基于流水数据及其关联的流水数据集合,以判断识别的准确性;

[0073] 第一数据存储模块用于将流水数据存储至缓存中;

[0074] 第二数据存储模块用于将综合存入关联性数据库中。

[0075] 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,该计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本发明所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括

随机存取存储器(RAM)或者外部高速缓冲存储器。

[0076] 作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双倍数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0077] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0078] 以上所述实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范围。因此,本发明专利的保护范围应以所附权利要求为准。

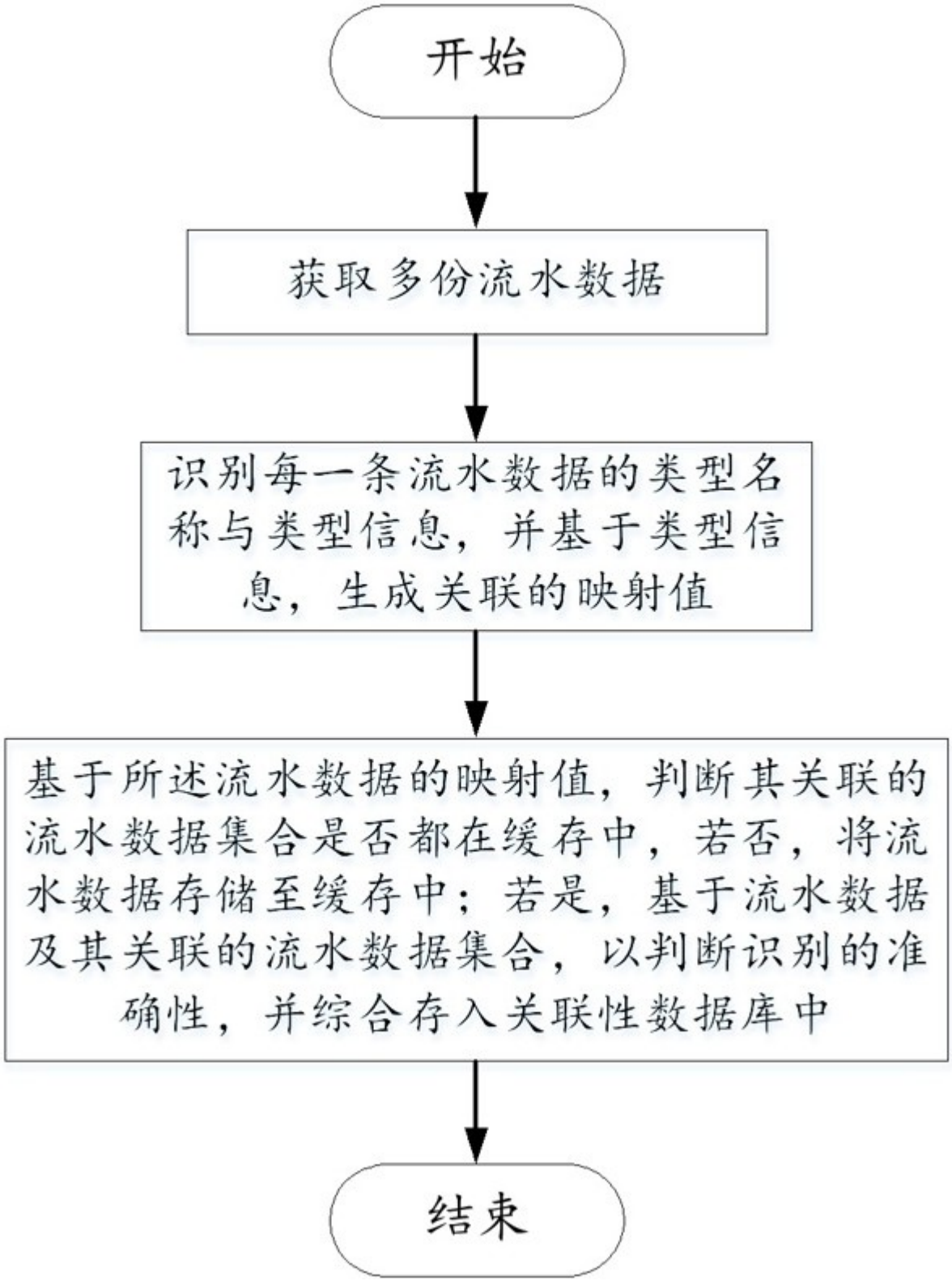

开始
获取多份流水数据
识别每一条流水数据的类型名称与类型信息，并基于类型信息，生成关联的映射值
基于所述流水数据的映射值，判断其关联的流水数据集合是否都在缓存中，若否，将流水数据存储至缓存中；若是，基于流水数据及其关联的流水数据集合，以判断识别的准确性，并综合存入关联性数据库中
结束

图1