文章编号：2095-2163(2024)02-0183-04 中图分类号：TP311.5 文献标志码：A

组态软件实时数据库的数据自动挖掘算法

段远志
（豫章师范学院 数学与计算机学院，南昌 330000）

摘 要：组态软件实时数据库中的数据存在缺失值、重复值等问题，导致在数据自动挖掘过程中所需时间较长、效率降低，因此提出组态软件实时数据库的数据自动挖掘算法。采集组态软件实时数据库内的数据，根据该数据结果，设定在数据挖掘的过程中项目内、外部效用的乘积为项目的事务效用，根据该事务效用提取数据的特征；将收集到的数据特征节点平移，设定坐标原点和数据挖掘的数量，得到自动挖掘数据信息的结果，完成组态软件实时数据库的数据自动挖掘研究。实验结果表明，组态软件实时数据库的数据自动挖掘算法的挖掘时间较短，提高了挖掘效率，证明了所提算法的有效性。

关键词：数据挖掘；数据库；组态软件；数据收集协议

Automatic data mining algorithm for real-time database of configuration software

DUAN Yuanzhi
（School of Mathematics and Computer Science, Yuzhang Normal University, Nanchang 330000, China）

Abstract: There are some problems in the real-time database of configuration software such as missing values and repeated values, which lead to long time and low efficiency in the process of automatic data mining. In this paper, an automatic data mining algorithm for real-time database of configuration software is proposed. The proposed method first collects data in a real-time database of configuration software and set the product of internal and external utility of a project as the transaction utility of the project in the process of data mining according to the data result and extract data features according to the transaction utility. The collected data feature nodes are translated, the coordinate origin and the number of data mining are set, and the results of automatic data mining are obtained. The research of automatic data mining in the real-time database of configuration software is completed. The experimental results show that the mining time of the automatic data mining algorithm in the real-time database of configuration software is shorter, which improves the mining efficiency and proves the effectiveness of the proposed algorithm.

Key words: data mining; database; configuration software; data collection protocol

0 引 言

如何高效地从海量数据中发现潜在的、有效的信息与规则，已成为关联规则挖掘领域的一个热点问题。实时数据库技术是自动化系统的核心技术，能够提供实时的数据管理和处理能力，为自动化系统的运行提供重要的支持。在自动控制组态软件中，实时数据库是一个非常重要的组成部分[1]，将数据挖掘算法自动加入实时数据库中，可以提供基本功能，包括数据存储、数据查询和数据访问控制。

蒋志龙[2]提出一种基于力控组态软件数据转储的实现算法，从一个嵌入式的实时数据系统出发，建立了一个实时数据查询处理系统，并以内存数据库为基础，构建出一个能够实现查询处理的关系系统。但该方法的数据挖掘只能依靠人工计算，并不能做到数据的自动化挖掘。王常武[3]提出改进的粒子群优化高效用项集挖掘算法，该算法应用于粒子群优化流程中，在当前代种群的高效用项集中以一定概率选择下一代种群的初始优化值。该算法增加了种群的多样性，使其能够挖掘更多的高效用项集。实验结果表明了该算法的有效性，但是在自动挖掘方面还存在挖掘效率较低的问题。因此，本文为改变上述方法的不足之处，提出组态软件实时数据库的数据自动挖掘算法。

1 组态软件实时数据库的数据自动挖掘

1.1 采集数据库内的数据

在组态软件中，要实现对数据的快速自动挖掘，

作者简介：段远志（1976-）男，硕士，讲师，主要研究方向：软件工程、数据挖掘、软件设计模式与体系结构。

收稿日期：2023-06-28

就必须解决数据的采集问题。想要做到这一点,就必须要有足够的数据储备。

数据收集指标是基于组态软件的实时数据库吞吐率和数据传输均衡的性能来设置的,这两个因素与数据传输协议有较强的相关性[4]。为此,基于数据快速收集网络的拓扑结构,采用数据收集协议来分配和收集数据,以提高数据收集指标的准确性和快速性。数据采集协议可以将数据中的信息迅速地分割成多帧,方便数据的传输,降低对系统空间资源的占用率[5]。在此基础上,通过对多个链路进行分组,从而避免由于链路拥塞导致的数据收集效率下降的问题。

将实时数据库内的数据进行初始化,设置迭代次数为 1,在任何一个种群中,粒子代表 1 个项集,计算保留下的数据,用数值大于阈值的项集表示,项集 $u(X)$ 的计算公式如下:

$$u(X) = T_c \times F_G \tag{1}$$

式中:T_c 代表数据库内的项集采集时间,F_G 代表数据库内的项集个数。

使用种群中的高效用项集,从当前群体的有效使用项集中选取下一代群体的最优初值[6]。以 s_i 表示在当前数据库的种群中第 i 个有效地使用项目集被选择的可能性,其计算公式如下:

$$S_j = \frac{F_{si} \times u(X)}{F_{sk} \times |f|} \tag{2}$$

式中:F_{si} 代表数据库种群中的高效用项集的数据,$|f|$ 代表集合 f 中的元素数量,F_{sk} 代表种群内的第 i 个用项集的适应数据。

若一个数据库上的节点上的第 j 个位置是 1,则表示与该位置相对应的项在一个项集上。如果不是,则表明与第 j 个位置相对应的项不存在于项集中。同时,通过在交易数据库中具有较高交易权重的交易项目集合的个数来确定各颗粒所对应的编码矢量的长度。为加快有效利用条目的挖掘速度,有必要对事务数据库中没有的条目进行裁剪[7-8]。

在每一次迭代过程中,计算效用值,确定有效的用项集,并选择下一次的群体,直至达到最大迭代次数,最终将输出所有有效的使用条目。当前微粒群的最优值并没有被直接保存给下一代微粒群,而是采用自动挖掘算法从当前微粒群的最优值中按一定概率选取[9]。该方法可以有效地改善群体的多样性,从而提高对有效使用项目的挖掘质量与效率。数据库内数据采集流程如图 1 所示。

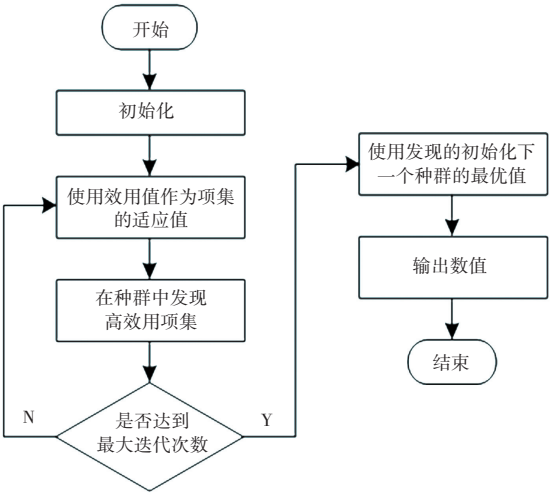

图 1 数据采集流程

Fig. 1 Flow chart of data acquisition

根据数据采集流程,将生成一个由微粒所表示的编码,调整输出数值的位置,从而得到存储在采集数据库中的数据[10],计算公式如下:

$$v_i = (v_{i1} + v_{i2} + v_{i3}) \times S_j \tag{3}$$

式中:v_{i1}、v_{i2} 和 v_{i3} 均表示随机变量。

为了保证所产生微粒的真实性,粒子 p_i 在编码矢量中的位置是通过逆码操作随机改变,确保粒子中的随机变量都经过预处理。

1.2 数据特征提取

组态实时数据库要求具有灵活的数据配置和清晰的数据层次结构[11]。为了实现对数据的快速自动挖掘,需要对数据的特性和操作过程进行分析[12-13]。首先利用协议对实时数据进行压缩和传输,减少在数据抽取时对网络存储空间的消耗;最后利用网络数据接入端口,对已有的数据进行分析,并针对不同的数据类型,选取合适的数据压缩和传输路径[14]。

表 1 为一个实时事务数据库的例子,该交易数据库含有 10 个不同的交易(T1、T2、…、T10)。交易记录 T1 中包括项目为(a,c,e),其内部效用值为(1、18、1);交易记录 T2 中包括项目为(b,d,e,f),其内部效用值为(7,1,1,1);以此类推,交易记录 T10 中包括项目为(b,c,e,f),其内部效用值为(6,2,2,4)。由表 2 可知,项目 A 到 F 的外在效用分别是(3,9,1,5,6,1)。其中,A 项所对应的外部效用是 1,即 A 项的重要度是 1;B 项所对应的外部效用是 9,即 B 项的重要度是 9,以此类推,F 项所对应的外部效用是 1,即 F 项的重要度是 1。

表 1　事务数据表

Table 1　Transaction data

交易记录	包含项目	项目内部效用
T1	(a,c,e)	(1,18,1)
T2	(b,d,e,f)	(7,1,1,1)
T3	(a,c,e)	(2,1,1)
T4	(d,e)	(1,1)
T5	(c,e)	(10,1,1)
T6	(b,f)	(4,2)
T7	(b,d,e)	(1,1)
T8	(a,c,d,e)	(3,25,3,1)
T9	(a,b,f)	(1,1,3)
T10	(b,c,e,f)	(6,2,2,4)

表 2 为数据的外部效用表。

表 2　外部效用表

Table 2　External utility

项 目	效用
A	3
B	9
C	1
D	5
E	6
F	1

设定在数据挖掘的过程中,项目的内、外部效用的乘积,即为项目的事务效用,计算公式如下:

$$U(w,T_c)=E_u(w)\times W_u(w,T_c)\times v_i \tag{4}$$

式中:w 代表事务的项目,$E_u(w)$ 为项目 w 的外部效用,$W_u(w,T_c)$ 表示事务项目 w 的内部效用。

由此可以证明,一个项集的效用函数不是单调性或反单调性的[15],即项目集合可以具有低效性、等效性和高效性 3 个子集。根据项目的事务效用,提取数据的特征结果,计算公式如下:

$$m_i=\lambda\times(f_{DY}+M_0)\times U(w,T_c) \tag{5}$$

式中:λ 表示指标叠加值,f_{DY} 表示通达性值,M_0 表示动态变化函数。

当数据特征确认后,使用者能够准确地检索和统计数据,通过向组态软件发送询问请求[16-17],实时数据库会作出相应的响应,同步地备份询问信息,并且将询问的记录保存下来,将待处理的数据直接存入实时数据库,并进行数据格式的变换,自动挖掘数据信息。

1.3　自动挖掘数据信息

在不断增加的交易数据中,可以发掘出潜在的敏感资讯,进而提示使用者需要关注的事项,以协助其作出决定。由于敏感数据往往在很短的时间内不能满足用户所设置的最低支撑,因而不容易被挖掘。因此提出了基于多层次同时处理和动态更新特征的组合敏感性指标,以实现对决策人员的实时敏感性[18]。

设置一个固定的时刻 t,作为一个敏感信息被挖掘的时间段[19]。通过对每个周期新增加的交易数据进行自动挖掘,形成一个数据库,并与以前的数据库进行对比和整合,得到一个周期中的数据与之前数据之间的变化关系,再根据各个时期所发生的信息变化程度,提出了两个敏感指标 d_{a_1}、d_{a_2},用于评价各个项目的敏感程度。在三维空间中,将收集到的数据节点特征平移,设定坐标原点为 o_0,数据挖掘的数量为 n,则自动挖掘数据信息的结果计算公式如下:

$$y=\frac{[(d_{a_1}+d_{a_2}-o_0)]}{m_i\times n} \tag{6}$$

针对测得的挖掘结果存在一定误差的问题,以便于融合海量实时数据的方法。为了实现这一目标,需要将各种数据的特征进行向量对比,并将特征与各数据的特征进行比较,以此确定资料在高维坐标系统中的相对位置,为数据的最后挖掘打下基础。随着网络技术和应用的不断发展,分布式数据库技术已经成为各企业使用的一种方式[20],而自动挖掘数据算法更适合于以此技术为基础,各个服务器在将各个数据库挖掘完之后,可以很容易地进行整合,这不仅可以提高挖掘的效率,还可以扩大其应用广度。

2　实验测试与分析

为证明本文提出的组态软件实时数据库的数据自动挖掘算法的有效性,将所提算法与基于力控组态软件数据转储的实现算法[2]、改进的粒子群优化高效用项集挖掘算法[3]进行了对比实验。此外,为验证所提算法的正确性,将所提算法运用在组态软件的实时数据库中,并对其能否满足规定的数据挖掘标准,能否取得良好的挖掘效果进行检验。

实验平台选择在 Matlab 环境中进行,采用 26 GHz 的 RF 数据处理芯片;组态软件的实时数据库工作电流损失设置为 32 mA,采集数据时的电流损失设置为 30 mA。在实验测试期间,根据实时数据库中的信道传输特性,对各节点间的平均距离进行了调整。采用了 6 核 4.4 GHz 的 AMDFX-8300;技嘉 990 X-D3P 主板,16 G 的内存;2 TB 的硬盘容量

(每分钟 7 200 转)。Eclipse-4.6.3 是一个综合的电脑语言程序设计环境,也是应用最广的开发环境之一。程序的最大运行时间设置为 60 min。

为了克服测试结果的单一性,以文献[2]和文献[3]中所提算法作为对照组,本文提出的方法作为实验组,对 3 种方法的性能进行综合评价。

实验采用 6 组公开的真实数据,实际数据组的相关信息见表 3。

表 3 实时事务数据库的相关特征

Table 3 Related characteristics of real-time transaction database

数据库名称	事务总数	总效用	项目总数
A	1 544 854	547 453	46 035
B	4 541	264 135	16 547
C	1 766	145 452	1 558
D	2 541	154 749	119
E	4 844	4 666 135	79
F	5 458	649 548	545

由于在不同的效用阈值下,所得到的有效用项集相同,因此在试验中对比了 3 种不同的挖掘方式所花费的时间,进而体现出 3 种方法的可实施性。

经过测试,3 种方法得出的数据挖掘时间的实验结果见表 4。

表 4 数据挖掘时间实验结果

Table 4 Experimental results of data mining time /ms

数据库	所提算法	文献[2]方法	文献[3]方法
A	15	29.3	32.1
B	14.9	19.6	29.1
C	15.7	22.1	22.6
D	15.7	25.9	26.78
E	12.8	20.6	20.4
F	13.5	19.9	19.5
平均值	14.6	22.9	25.08

由表 4 可知,采用本文算法进行 6 组不同的数据库的数据挖掘,其挖掘时间均小于另外两种方法,节省时间分别为 8.3 ms、10.48 ms,提高了挖掘效率,证明了所提算法的可行性。

3 结束语

本文使用组态软件进行信息快速数据定位算法,能够高效地、迅速地收集信息资料,并对带有特征的信息进行分类,进而实现信息的有效挖掘。实验结果表明,应用数据采集协议能够定期的从数据库间接提取信息,同时方便信息的对外传送,因此有效减少了信息收集能耗,减少了对空间数据的浪费。本文通过改进算法的实时性,对其进行进一步的优化。然而,在实际应用该算法时,需要考虑其优缺点,同时注意数据质量和样本大小等因素,以确保算法的可靠性和准确性。

参考文献

[1] 杨军庄,文德,张育辉. 基于数据挖掘的电能误差数据自动化校正方法[J]. 电子设计工程, 2022, 30(24):153-156.

[2] 蒋志龙. 一种基于力控组态软件数据转储的实现[J]. 能源技术与管理,2021,46(4):165-166,182.

[3] 王常武,尹松林,刘文远,等. HUIM-IPSO:一个改进的粒子群优化高效用项集挖掘算法[J]. 小型微型计算机系统,2020,41(5):1084-1090.

[4] 赵永国. 基于数据挖掘算法的智慧档案馆海量数据快速提取方法[J]. 自动化技术与应用,2022,41(9):40-43,186.

[5] 鹿有杰,任有志,胡涛. 基于组态软件自动化立体仓库管理系统的开发[J]. 科技资讯,2022,20(16):27-29.

[6] 李翀,李兵,张秀丽,等. 基于数据挖掘的电力计量设备生产计划与排程算法研究[J]. 制造业自动化,2023,45(4):88-91.

[7] 蔡勋玮,赵俊,马丛,等. 新供电服务指挥系统诉求响应数据自动挖掘方法[J]. 自动化技术与应用, 2022, 41(7):62-68.

[8] 王景兰,方晓. 基于层次频繁模式树的数据自动挖掘算法[J]. 上海电机学院学报, 2022, 25(4):239-242,248.

[9] 郭娟,杨为民,石亚和. 基于微粒群算法的二维最大熵图像分割方法[J]. 计算机仿真, 2005, 22(11):94-97.

[10] 吴浩,李廷鹃,李玉健,等. 基于数据挖掘的数据库信息查询访问控制系统设计[J]. 电子设计工程, 2023, 31(18):6-9.

[11] 胡自松,王丽珍, Tran V ,等. 基于图数据库的空间频繁并置模式挖掘[J]. 计算机科学与探索, 2022,16(4):806-821.

[12] 高源,施伟谊,周亦华,等. 一种面向动态不平衡数据流的集成超限学习机分类算法[J]. 复旦学报(自然科学版), 2023, 62(3):352-361.

[13] 黄蔓云,卫志农,孙国强,等. 数据挖掘在配电网态势感知中的应用:模型,算法和挑战[J]. 中国电机工程学报, 2022, 42(18):6588-6598.

[14] 刘利民,张勇. 多标签隐性知识显性化下的数据挖掘算法[J]. 计算机仿真,2023,40(4):504-508.

[15] 周江,陈扬,虞凌云. 基于效用函数的无源光网络动态带宽分配算法[J]. 计算机与现代化, 2023(4):106-110.

[16] 俞冠中,田青旺,彭飞,等. 基于国密算法的核电站 DCS 逻辑组态软件信息安全防护方法研究[J]. 制造业自动化, 2023, 45(12):61-64.

[17] 朱文,邓大为,吴争荣,等. 基于图形组态软件的新能源变电站图模一体化生成方法[J]. 科学技术与工程, 2023, 23(24):10372-10378.

[18] 陶文伟,吴金宇,江泽铭,等. 基于数据挖掘的智能电网故障处置辅助决策系统[J]. 测试技术学报,2023,37(2):127-134.

[19] 黄诚,赵倩锐. 基于语言模型词嵌入和注意力机制的敏感信息检测方法[J]. 计算机应用, 2022, 42(7):2009-2014.

[20] 刘勇. 信息管理中计算机数据库技术的应用探讨[J]. 中国宽带, 2022, 18(11):28-30.