第32卷 第20期 计 算 机 工 程 2006年10月
Vol.32 № 20 Computer Engineering October 2006
·软件技术与数据库· 文章编号：1000—3428(2006)20—0067—02 文献标识码：A 中图分类号：TP311.52

双时态周期数据模型的研究

任家东，龚 冰
(燕山大学信息科学与工程学院，秦皇岛 066004)

摘 要：在经典的时态数据模型HMAP的基础上，通过增加事务时间参数*delay*的方式，提出了一种支持多时间粒度，可以有效处理具有周期特性数据的双时态数据模型PMB，该数据模型可以有效地解决HMAP模型存储中记录条目繁多和时间交叉的问题，从而节约存储空间，提高数据的查询速度。 理论分析和实验结果表明PMB数据模型可以把需要存储为*n*条记录的数据简化为有限的几条，而且当*n*值很大，甚至无穷的时候，这种存储方式具有明显的优势。
关键词：双时态；周期事务；固定延迟

Research on Bi-temporal Periodic Data Model

REN Jiadong, GONG Bing
(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

【Abstract】 This paper proposes a new multi-granularities bi-temporal data model PMB, based on the classical HMAP model, by introducing the definition of transaction period parameter *delay*, to process the periodic data. The proposed model can efficiently deal with the problems of enormous record items and intersectant time points existing in the HMAP model, which will save the store space and improve the speed of searching for data. Theoretical analysis and experimental results show that the PMB model can make *n* items of the list cut short to a few items. When the value of *n* is very enormous, the advantage of this model is very prominent.
【Key words】 Bi-temporal; Periodic transaction; Fixed delay

双时态数据模型是指同时支持两种类型时间(有效时间和事务时间)的数据模型。有效时间是指事务对象在现实世界中为真的时间，事务时间是指此对象记录作为数据存储在数据库中的时间[1]。

医疗信息系统(Hospital Information System，HIS)是时态数据被广泛应用的领域[2]，其中的临床医学领域存在着大量与时间相关的信息[3]。临床医学中的病史记录是HIS的重要资源，其数据类型非常丰富。为了不丢失数据中的信息，数据模型往往采用较细的或可变的时间粒度表示。

在时态数据中存在着大量具有周期属性的数据。对于周期问题，目前的研究大多集中在周期规律的数据挖掘方面[4]，而很少涉及数据的处理。现有的数据模型很少关注周期数据的存储等问题。在普通的时态数据模型中，周期事务通常被当作重复发生的非周期事务处理，这势必造成资源的浪费。另外，现有的时态数据查询语言大多不支持周期查询，TSQL2标准[5]也不支持周期查询。

本文首先介绍了一种时态数据模型HMAP[6]，同时提出了一些HIS中典型的病史记录实例。HMAP模型可以基本表达清楚实例中的病史记录。但实例中含有大量的周期数据，如果用HMAP概念下的双时态数据模型表示，那么记录的条数将十分庞大，因此本文提出了一种改进的数据模型PMB。PMB模型可以较好地处理具有周期特性的数据。对于大量涉及周期数据的领域，周期数据出现的越频繁，PMB的优势就越明显。

1 时态数据模型 HMAP

1.1 HMAP 模型

普通模型通常采用[*start*, *end*]的格式表示时间，且*start*、*end*仅为一个时间点。由Carlo Combi和Giuseppe Pozzi于1998年提出的HMAP模型[6]在此基础上增加了参数*duration*，加强了模型表达时态信息的能力。HMAP模型是一种将自然语言数字化的模型，在这一方面，以前的模型大多不能较好地表达自然语言中比较复杂的时间信息。

HMAP是一种时态数据模型，重点研究的是双时态数据中有效时间的表达，它采用 [*start*, *duration*, *end*] 的格式表示有效时间。而且，HMAP模型允许用户采用不同的时间粒度、不确定度表示每个参数，以此为基础的双时态数据模型在表达能力上得到了很大的提高。

1.2 病史记录的实例

在HIS中，病史资料的记录非常重要。然而，存储和查询病史记录一直是HIS发展的“瓶颈”问题。现在给出一些在本文中被多次使用的医疗信息记录。

记录 A 病人的一次腹痛发生的时间大约是从2004年7月25日下午的1:00～1:10开始，直到3:30～3:40左右才结束，持续时间大约为2h25min～2h40min；

记录 B 从2004年7月27日开始，每隔一天对病人进行一次防凝血治疗，时间大约为上午9:30～11:00之间，持续80min，一共持续了两周；

记录 C 从2004年8月1日开始，每周的周二、周五分别对病人进行一次化疗，时间大约为下午2:30～4:00之间，持续70min。

以上记录在病史资料中具有很强的代表性，但普通数据模型很难处理这些记录。

基金项目：河北省博士基金资助项目(B200322)
作者简介：任家东(1967－)，男，博士、教授，主研方向：数据库技术；龚 冰，硕士生
收稿日期：2006-01-03 **E-mail：**gongbing_2005@yahoo.com.cn

1.3 HMAP 的实例表达

设 HMAP 的双时态数据结构为[*v-start, duration, v-end, t-start, t-end*]。其中，*v-start* 为有效时间的开始时间，*duration* 为有效时间内事务对象实际持续的时间长度，*v-end* 为有效时间的结束时间，*t-start* 为事务时间的开始时间，*t-end* 为事务时间的结束时间。

设时间粒度为(y，m，d，h和m_i分别代表年、月、日、小时和分钟，单位为位)：

v-start: yyyy / mm / dd / hh / m_im_i

v-end: yyyy / mm / dd / hh / m_im_i

t-start: yyyy / mm / dd

t-end: yyyy / mm / dd

duration: hh / m_im_i

UC 表示不确定的时间，采用 24h 制表示时刻，病史记录见表 1，从表 1 中可发现 HMAP 存在一些缺陷。

表 1 基于 HMAP 的病史记录

编号	记录名	*v-start*	*duration*	*v-end*	*t-start*	*t-end*
1	A	2004/07/25/13/00, 2004/07/25/13/10	02/25, 02/40	2004/07/25/15/30, 2004/07/25/15/40	2004/07/25	*UC*
2	B	2004/07/27/09/30	01/20	2004/07/27/11/00	2004/07/27	2004/07/28
3	B	2004/07/29/09/30	01/20	2004/07/29/11/00	2004/07/29	2004/07/30
…	…	…	…	…	…	…
9	B	2004/08/10/09/30	01/20	2004/08/10/11/00	2004/08/10	*UC*
10	C	2004/08/03/14/30	01/10	2004/08/03/16/00	2004/08/03	2004/08/09
11	C	2004/08/06/14/30	01/10	2004/08/06/16/00	2004/08/06	2004/08/12
12	C	2004/08/10/14/30	01/10	2004/08/10/16/00	2004/08/10	2004/08/16
13	C	2004/08/13/14/30	01/10	2004/08/13/16/00	2004/08/13	2004/08/19
…	…	…	…	…	…	…
10 +*n*	D	…	…	…	…	…

首先，它在处理病史记录 B 和记录 C 时，记录数目繁多。其次，在表述病史记录 C 时，由于事务时间出现了周期交叉，因此可能导致错误的操作甚至死循环。

2 周期变粒度双时态数据模型 PMB

基于 HMAP 模型本文提出了一种改进的模型 PMB(Perio-dic Multi-granularities Bi-temporal Data Model)，它可以有效地解决 HMAP 中记录条目繁多和时间交叉的问题。

2.1 双时态结构说明

在 HMAP 模型基础上，增加参数 *delay* 表示相邻事务操作之间的时间间隔。

定义 1 周期事务是指处理周期数据时，重复进行的某一事务操作。每一次这种事务操作称为一个周期事务(periodic transaction)。相邻的两次周期事务操作之间的时间间隔称为一个事务周期(transaction period)。

定义 2 PMB 模型的双时态结构为

[*v-start, duration, v-end, t-start, delay, t-end*]

其中，*v-start*、*duration*、*v-end* 和 *t-start* 的含义与 HMAP 模型完全相同。*delay* 表示相邻事务操作之间的时间延迟。对于非周期事务 *delay* 取 0 值，对于周期事务 *delay* 取事务周期值，同时 *delay* 也作为判定周期数据的标志。*t-end* 表示事务时间的结束。

2.2 PMB 的实例表达

PMB 中相关参数的设定与 HMAP 对应的设定相同。另外，参数 *delay* 的时间粒度设为“ *delay:* dd ”。基于 PMB 模型的病史记录参见表 2。PMB 将需要存储为 *n* 条记录的数据集简化成有限的几条。当 *n* 值较小时，这种表示方法的优势不是很明显，但当 *n* 值很大，甚至无穷时，它的优势就变得非常明显。周期数据的许多信息是重复出现的，因为有效时间和事务时间往往存在着一定的关系，所以可优化存储方式。

表 2 基于 PMB 的病史记录

编号	记录名	*v-start*	*duration*	*v-end*	*t-start*	*delay*	*t-end*
1	A	2004/07/25/13/00,20 04/07/25/13/10	02/25, 02/40	2004/07/25/15/30 2004/07/25/15/40	2004/07/25	0	*UC*
2	B	09/30	01/20	11/00	2004/07/27	02	2004/08/10
3	C	14/30	01/10	16/00	2004/08/03	07	*UC*
4	C	14/30	01/10	16/00	2004/08/06	07	*UC*
5	D	…	…	…	…	…	…

以病史记录C为例，由于治疗每周都进行两次，且时间均为“ 下午 2:30～4:00，持续 70min ”，虽然最小时间粒度要取到分钟，但仅用“ hh/m_im_i ”的形式即可，而不必一定写成“ yyyy/mm/dd/hh/m_im_i ”的形式。在这一方面，非周期数据同样可以适用。不过，PMB模型仍采用普通方式处理非周期数据，这主要是基于模型兼容性的考虑。另外，PMB模型采用单独处理相邻事务时间范围内的周期数据的方式，较好地避免了HMAP模型中时间交叉的情况。

2.3 性能分析

普通的数据模型在处理病史记录时，常用的一种方法是用文字进行描述，这种方法虽然有较高的处理效率，但是查询起来往往受到诸多限制。

HMAP 模型很好地解决了病史记录 A 这类情况。但是在 HIS 中，除了表述复杂的数据以外，还有许多周期数据。HMAP 模型表达信息的能力较强，但记录比较冗长。如病史记录 C，HMAP 模型几乎无法处理。

下面重点分析数据的时态部分存储。由于 PMB 模型存储变粒度数据比存储固定粒度数据所需的空间还要少，因此此处仅以存储固定粒度的情况为例说明。

证明 设x为变量，A_i为数据，$C(x)$为变量x所占的存储空间，$C(A_i)$表示数据A_i所占的存储空间，数据集S中包含k种非周期数据、m种周期数据，n_i表示周期数据A_i发生次数，n表示集合$[n_1,\ldots,n_m]$中的最小值。则$n\geq 2$，且数据集S中周期数据与非周期数据的种类数之比可以用m/k表示，令m/k的最小取值为Min(m/k)。

设HMAP的双时态数据结构为结构I，PMB的双时态数据结构为结构II。令$C_I(S)$和$C_{II}(S)$为数据集S分别按照结构I和结构II存储时所占的空间，那么式(1)成立，可以推知：数据集S按照结构II比按照结构I存储所占的空间要小。

$$C_I(S) > C_{II}(S) \tag{1}$$

$$m/k > 1/(22\times n-23) \tag{2}$$

由以上命题假设，可以得到推理如下：

因为

$C(v\text{-}start) = C(v\text{-}end) = 12$

$C(t\text{-}start) = C(t\text{-}end) = 8$

$C(duration) = 4$，$C(delay) = 2$

所以

$C_I(A_i) = 44$

$C_{II}(A_i) = 46$

因为

$C_I(S) \geq (m\times n+k)\times 44$

$C_{II}(S) = (m+k)\times 46$

所以

$$(m\times n+k)\times 44 > (m+k)\times 46 \Rightarrow C_I(S) > C_{II}(S)$$

因为$(m\times n+k)\times 44 > (m+k)\times 46 \Leftrightarrow m/k > 1/(22\times n-23)$，所以要使式(1)成立，只需式(2)成立。其中，$n$ 值和 Min(m/k)的关系见表 3，证毕。

(下转第 94 页)