基于时态层次链的 Data Cube 多版本维护方案①

肖 磊 胡众义 (温州大学 物理与电子信息工程学院 浙江 温州 325035)

摘 要: 针对 Data Cube 的模式变动造成的多版本问题，对时态聚集关系与时态层次链进行了形式化描述，并基于这些关系实现了多个版本的 Data Cube 的统一生成算法，不仅可以高效地进行 Data Cube 多个版本的维护，而且在 OLAP 查询也可以基于时态层次链来执行，从而提高系统的整体效率。

关键词: 时态层次链 多维数据集 多版本维护

A Multi-Version Maintain Scheme of Data Cube Based on Temporal Hierarchy Chains

XIAO Lei, HU Zhong-Yi

(College of Physics and Electronic Information Engineering,Wenzhou University, Wenzhou 325035, China)

Abstract: This article first gives formalization definitions of temporal aggregation relation and temporal hierarchy chain for solving the data cube's multi-version problem caused by pattern's changes. Then a uniform algorithm for generating the multi-version data cube is proposed. It cannot only maintain the multi-version data cube effectively but also optimize the OLAP queries based on temporal hierarchy chains and thus the efficiency of system is improved.

Keywords: temporal hierarchy chains; data cube; multi-version maintain

随着数据库技术的广泛应用和发展，产生了数据仓库、联机分析处理等一系列新技术，并且在实践中得以逐步应用。在 OLAP(OnLine Analytical Processing)中为了提高查询响应时间，采用了 data cube[1]预计算技术,从而有效的提高了 OLAP 的效率。

1 引言

1.1 文章安排

本文第 2 节对时态层次链进行了形式化描述，第 3 节对基于时态层次链的 Data Cube 生成算法与维护操作进行了讨论，第 4 节通过实验分析了相关算法的性能，第 5 节对全文进行了总结。

1.1.1 基本介绍

在现在的对 Data Cube 的研究中，都集中于研究如何有效的对其进行有效的计算，或者是对 Data Cube 进行压缩存储的算法，在这些研究过程中，都是假定 Data Cube 的模式是稳定的，即 Data Cube 中的维、层次及它们之间的关系都是不变化的。但在实际应用中，往往会出现 Data Cube 的模式发生变动的情况，即多版本 Data Cube 的问题，例如：产品的分类变化会导致产品维的层次聚集关系发现变化，现实世界中区域的变动(如香港的回归等)会导致地区维的变动等。文献[2]提出了一种多版本数据仓库联合查询的算法，文献[3]则研究了多版本数据仓库中的索引方案，文献[4]对 OLAP 查询中的时态相关查询进行了分析，但并没有涉及到多版本 Data Cube。

基于上述的分析，我们提出了 Data Cube 中的时间层次链，并进行了形式化的描述，用于解决 Data Cube 的多版本问题。

2 相关概念与定义

多版本 Data Cube 的维护，实质就是要对各个时期的 Data Cube 的模式进行记录，在进行聚集操作或

① 收稿时间:2009-04-13

查询时能根据相关的时间段来进行相应的操作，从而在 Data Cube 有多个版本时，用户可以不用考虑不同版本 Data Cube 的差异，即用户可以“透明”地进行操作。为了更准确的描述 Data Cube 的版本特征，我们提出了时态层次链的概念。

定义 1. (层次 H)层次 H 可定义为一个某个域中值的集合。并定义一个仅包含一个值 all 的层次 ALL，表示维中的最高层次。

定义 2. (层次之间的聚集关系 $\preceq$)设有两个层次 H',H，集合 PH 为 H 的一个划分。如果存在函数 f，使得任取 $ph \in PH$，均有 $f(ph) \in H'$。则称层次 H 能聚集运算到 H'，记为 $H' \preceq H$。

定理 1. 对于任意一个层次 H，均有 $ALL \preceq H$。证明略。

定义 3. (层次之间的直接聚集关系 $\prec$)在若干个层次的集合中，对于两个层次 H', H，如果有 $H' \preceq H$，并且在这个集合中，不存在另外一个层次 H''，使得 $H' \preceq H'' \preceq H$，则称层次 H 直接聚集到 H'，记为 $H' \prec H$。

定义 4. (维)Data Cube 中的维可描述为一个二元组 $<HS, \prec_S>$，其中 HS 为层次的集合，$\prec_S$ 为这些层次间直接聚集关系 $\prec$ 的集合。

上述的定义采用形式化的描述方法严格定义层次之间的聚集关系与维，从而在进行 Data Cube 计算时，能够根据层次之间的 $\prec$ 关系来进行计算维与 Data Cube 的聚集计算。但是如果出现 Data Cube 的模式发生变动的情况，则无法根据变动的情况来更改相关的聚集计算方式，为了解决这一问题，我们引入了时态聚集关系 $\prec_t$。

定义 5. (层次之间的时态聚集关系 $\prec_t$)时态聚集关系 $\prec_t$ 可描述为一个二元组 $<\prec, Period>$，其中 $\prec$ 如定义 3 中所示，$Period$ 为此直接聚集关系的有效期，包括：$StartT$：开始时刻，$EndT$：结束时刻。$EndT$ 中包括一个变量：uc，表示“直到当前”的时间概念。

由定义 5 可知，时态聚集关系是对直接聚集关系的扩展，即在直接聚集关系的基础上增加时间有效期的描述，两个层次之间可以有多个层次之间的时态聚集关系，从而能够记录 Data Cube 模式的多个版本，在进行聚集计算时可以根据相关的 $\prec_t$ 关系来进行区分。

定理 2. 如果两个层次间有多个时态聚集关系：$\prec_t^1$，$\prec_t^2$，$\prec_t^3$，……，$\prec_t^m$，则任取其中的两个时态聚集关系 $\prec_t^i$，$\prec_t^j$，均有 $\prec_t^i .Period \bigcap \prec_t^j .Period = \Phi$。证明略。

为了更明确的描述基于时态聚集关系的聚集计算过程，我们定义了时态层次链：

定义 6. (基于时态聚集关系的维)Data Cube 中的维可描述为一个二元组 $<HS, \prec_{tS}>$，其中 HS 为层次的集合，$\prec_{tS}$ 为这些层次间时态聚集关系 $\prec_t$ 的集合。

定义 7. (时态层次链)设 D 为 Data Cube 中一个基于时态聚集关系的维，所包含的层次集合为 HS，对于若干个 HS 中的层次 $h_1, h_2, \cdots, h_n$，如果存在 $n-1$ 个层次之间的时态聚集关系，使得 $h_1 \prec_{t1} h_2 \prec_{t2} \cdots \prec t_{n-1} h_n$ 成立，同时有 $\bigcap_{i=1}^{n-1} \prec_{ti} .Period != \Phi$，则称 $h_1, h_2, \cdots, h_n$ 为维 D 中的时态层次链。$\bigcap_{i=1}^{n-1} \prec_{ti} .Period$ 称为此时态层次链的有效时间，记为 $HS.Period$。

定义 8. (极大时态层次链)设层次 $h_1, h_2, \cdots, h_n$ 为维 D 中的一个时态层次链，如果不存在层次 h_0，使得 $h_0 \prec_{t0} h_1 \prec_{t1} h_2 \prec_{t2} \cdots \prec t_{n-1} h_n$ 或 $h_1 \prec_{t1} h_2 \prec_{t2} \cdots \prec t_{n-1} h_n \prec_{t0} h_0$ 成立，则称 $h_1, h_2, \cdots, h_n$ 为维 D 中的一个极大时态层次链。

定理 3. 如果 $h_0 \prec_{t0} h_1 \prec_{t1} h_2 \prec_{t2} \cdots \prec t_{n-1} h_n$ 为维 D 中的一个极大时态层次链，则有 $h_0 = ALL$。证明略。

3 基于时态层次链的Data Cube生成与相关操作

3.1 基于时态层次链的 Data Cube 生成算法

上述的定义在对层次及层次之间聚集关系进行形式化描述的基础上，通过引入有效期的概念，对时态聚集关系及时态层次链等概念都进行了严格的定义，从而可以在生成 Data Cube 的时候，根据单元格的时间维信息，来确定聚集的路径，并生成聚集值。基于时态层次链的 Data Cube 算法如下所示：

算法 1 基于时态层次链的 Data Cube 生成

输入：带有时间维信息的基表元组 T，Data Cube 中的元数据信息，Data Cube

输出：包含有基表元组 T 各聚集值的新 Data Cube

步骤：

① 从元数据信息中，得到每个维上的极大时态层次链个数 $num_i(1 \le i \le d)$，d 为维的个数；

② 获得基表元组 T 的时间维信息 $time_T$；

③ for $k=1$ to $\prod_{i=1}^{d} num_i$ //每个维上选择一个极大时态层次链，一共有 $\prod_{i=1}^{d} num_i$ 种组合

④ {

⑤ 生成一个各维上极大时态层次链的组合 $<tChain_1, tChain_2, \cdots, tChain_d>$；

⑥ 如果对于每一个极大时态层次链 $tChain_i$，均有 $time_T \subseteq tChain_i.Period$，则：

⑦ {

⑧计算基表元组 T 基于各个维极大时态层次链直到 $<all, all, \cdots, all>$ 的聚集值，并将所

⑨ 有的聚集值加入到原 Data Cube 中；

⑩ }

⑪ }

算法的主要思路是对于将要进行聚集计算基表元组，都去检测其时间是否在每个维上各条极大时态层次链的有效时间内，如果是的话，则可以采用各种 Data Cube 的聚集算法来生成聚集值，这样当 Data Cube 的模式发生变化时，可以直接在原有的 data cube 中进行操作，而不用再建新的 Data Cube。另一方面，如果 Data Cube 有 n 个维，每个维上有 m 个极大时态层次链，则算法复杂度为 $O(nm)$。

3.2 基于时态层次链的 Data Cube 模式维护操作

以下是对 Data Cube 的各种模式变化时，所需要进行的维护操作：

(1) 层次中的聚集关系发生变动。这也是最经常发生的模式变动，例如，公司 2008 年以前在华中地发起的某个省份基本没有业务，因此在聚集时，这个省的销售额是跟另外几个省一起聚集到上一个层次“区域”的“其他”中，而 2008 年后业务开展起来了，决策者要求这个省份的业务额向上聚集到“区域”的“华中”中。一般来讲，如果层次中的聚集关系在某一时刻 t 由 $\prec$ 变为 $\prec'$，则层次间的时态聚集关系变化如下：

① 时态聚集关系 $<\prec,(t_b,uc)>$ 修改为：$<\prec,(t_b,t)>$；

② 层次间增加一个时态聚集关系 $<\prec,(t,uc)>$。

(2) 删除一个层次。在删除维中的某个层次 h 时，其维护操作如下：

① 找出所有的包含层次 h 的极大时态层次链 $chain_i$

② 对于每个极大时态层次链 $chain_i$，如果层次 h 为最底层的层次，则直接删除；否则

③ 找到层次 h 在此极大时态层次链中的上一层次 ha 与下一层次 hb

④ 层次 ha 与层次 hb 间增加一个时态聚集关系 $<\prec',(t,uc)>$，其中时刻 t 为进行删除操作的时间，$\prec'$ 中的聚集函数 f 由 hb 到 h 的聚集函数 fb 与 h 到 ha 的聚集函数复合生成。同时层次 ha 与层次 h，层次 h 到层次 hb 之间的时态聚集关系的有效期中的 uc 均修改为 t。

(3) 增加一个层次。假设是需要在维中层次 ha 与层次 hb 中增加一个层次 h，则维护操作如下：

① 层次 ha 与层次 hb 之间的时态聚集关系的有效期中的 uc 修改为 t，其中时刻 t 为进行更新操作的时间。

② 增加层次 ha 与层次 h，层次 h 到层次 hb 之间的相应时态聚集关系。

如果还有其他的模式变动，则都可以映射成对时态聚集关系的修改。因此，在 Data Cube 中引入时态层次链后，当 Data Cube 的模式发生变化时，系统能够以统一的观点来对处理 Data Cube 模式的各个版本，同时在聚集计算时，只需要调用算法 1 来进行各个基本元组的聚集值，而不用去考虑各个版本之间数据的差异。

4 实验结果与性能分析

为了验证相关算法的有效性，我们对算法进行了测试。测试所采用的数据集为 TPC-R[5]，实验机器为一台 Intel Pentium IV 2.6GHZ，512M 内存，运行 Windows 2000 Server 的 PC 机。

实验一比较了引入了时态层次链的概念后，各种 Data Cube 生成算法 BUC[6]，PetaCube[7]，QC-Tree[8]中 DFS 及 SDC[9]的执行时间比较，其结果如图 1 所示。由图可以看出，无论对于何种算法，加入时态层次链后，所引起的时间耗费不会超过原有执行时间的 3%。

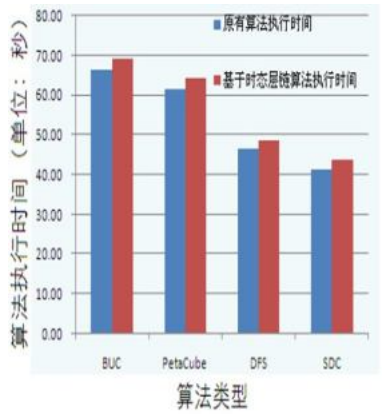

图 1　时态层次链对 Data Cube 生成算法的影响

实验二是考察引入时态层次链后，对于 OLAP 查询优化作用。我们随机选取了多个 OLAP 查询，并考察经过多次模式变化后查询时间的变化，同时与普通 Data Cube 上的查询时间相比较，其结果如图 2 所示。由图可以看出，对于普通的 Data Cube，每一次模式变化，都要导致 Data Cube 的重构与分解，在查询时也需要对 Data Cube 的多个版本进行联合查询才能得到查询结果，而基于时态层次链的 Data Cube 则可以用统一的视图来看待和管理 Data Cube 的多个版本，因此执行效率会更高，并且查询越多，模式变化的次数越多，优化效果就越为明显。

图 2　时态层次链 Data Cube 的优化效果

5　总结

本文对时态层次聚集关系进行形式化描述，并定义了时态层次链与极大时态层次链，针对各种模式变化的情况，分析了时态层次链保存 Data Cube 多版本信息的方式，从而减少查询耗费的代价，改善了系统的性能。以后的研究方向将继续对极大时态层次链的维护操作进行研究，并对 Data Cube 生成算法进行改进，以进一步提高系统的效率。

参考文献

1 Gray J, Bosworth A, Layman A, Pirahesh H. Data Cube:A Relational Aggregation Operator Generalizing Group-By,Cross-Tab,and Sub-Totals.IEEE Inf.Conf. Data Zngineering, New Orleans, Louisiana, 1996, 152－159.

2 Rizzi S, Golfarelli M. X-time: schema versioning and cross-version querying in data warehouses. Proc. of the International Conference on Data Engineering (ICDE). 2007.

3 Jouini K, Jomier G. Indexing multiversion databases. Proc. of ACM Conference on Information and Knowledge Management (CIKM). 2007.

4 Mendelzon AO, Vaisman AA. Temporal queries in OLAP. Proc. of the International Conference on Very Large Data Bases (VLDB). 2000.

5 Transaction Processing Performance Council TPC. TPC benchmarks H and R (decision support)(OL). Standard Specification,Transaction Processing Performance Council (TPC).October 1999. http:// www. tpc.org/

6 Beyer K, Ramakrishnan R. Bottom-up computation of sparse and iceberg CUBEs. Proc. of the ACM SIG MOD Int'1 Conf. on Management of Data. ACM Press, 1999.

7 Sismanis Y, Deligiannakis A, Roussopoulos N, et al. Dwarf: Shrinking the Peta Cube. Proc. of ACM SIGMOD Int'1 Conf. on Management of Data. Madison, USA: ACM Press, 2002.

8 Lakshmanan L, Pei Jian, Zhao Yan. QC-trees: An efficient summary structure for semantic OLAP. Proc. of ACM SIGMOD International Conference on Management of Data. New York, USA, 2003.

9 师智斌,黄厚宽,刘红敏.一种保持语义的压缩数据立方体结构.计算机工程, 2008,34(13):37－39.