

基于粗集理论的视图分层更新方法^{*})

李文海 冯玉才 马晓鸣 尹建章

(华中科技大学计算机学院 数据库与多媒体技术研究所 武汉 430074)

摘要 为缩短查询响应时间和提高决策支持能力,数据仓库中存储了大量的实化视图,视图更新成为影响系统整体性能的一个重要因素。本文分析了现有方法处理大规模视图更新的不足,提出一种基于粗集约简发现中间视图依赖层次的方法,利用有效代价确定更新次序以达到更新费用最低,随后分析算法的复杂性并通过实验证明其有效性。

关键词 数据仓库,实化视图,层次更新,粗集,约简

Hierarchical Updating Method Based on Rough Sets Theory in View Maintenance

LI Wen-Hai FENG Yu-Cai MA Xiao-Ming YIN Jian-Zhang

(Institute of Database and Multi-media Technology, Huazhong University of Sci. & Tech., Wuhan 430074)

Abstract For the sake of acquiring the quick response and enhancing the power of the decision support, data warehouses maintain large numbers of materialized views. And view updating becomes a significant factor of the overall performance. This paper analyzes the defect of the existing methods in handling the updating of the massive views, and proposes a new method based on rough sets theory to discover the correlative hierarchy of the mutual views. The valid cost is utilized to acquire the updating order with the minimal expense. The validity and complexity of the algorithm are given in the end.

对于任意查询 Q 和源关系 R_i , 若 R_i 中两列 $R_i.C_a$ 和 $R_i.C_b$ 均出现在 Q 的谓词 $P(Q)$ 中, 可在 $\mathcal{R}$ 中用 R_{a_i} 和 R_{b_i} 替代 R_i , 故下文假设 $\mathcal{R}$ 中各关系在 $\mathcal{Q}$ 的任意查询下均只有一个属性被访问。DM4 选择性地使用包或集语法处理集函数且将其上推到连接后执行, 我们在设计 DMWS 的视图选择和更新时主要面向 SPJ 查询。给定源关系集 $\mathcal{R}$, 一个 SPJ 视图 V :

$$\Pi_{proj \sigma and} (R_1 \bowtie \cdots \bowtie R_i \bowtie \cdots \bowtie R_n) \quad (1)$$

存储了在关系 $(R_1, \dots, R_i, \dots, R_n) \subseteq \mathcal{R}$ 上执行连接 $\bowtie$ 、选择 σ 和投影 Π 后的元组结果集。若发生变更 ΔR_i , 仅当 R_i 在投影或选择谓词的对象列表中时才能触发视图 V 的更新^[2], 且此时更新代价仅包含从数据源 R_i 到视图维护器传输 ΔR_i 的代价。我们在监视器中首先用列范围规则判断对 ΔR_i 过滤, 仅将符合规则的 ΔR_i 发送到维护器。列范围规则由一组形如 $RP(R_i) = [(R_i.C_a \text{ op } c_i^a) \vee (R_i.C_a \text{ op } c_i^b) \vee \dots] \vee [(R_i.C_b \text{ op } c_i^1) \vee \dots]$ 的析取式组成; op 表示操作符“<”, “<=”, “=”, “>”, “>=”; 单查询对 R_i 中属性 C_a 的选择约束为 $R_i.C_a \text{ op } c_i^a$ 且 $c_i^a \neq \emptyset$, 满足 $[(R_i.C_a$

个层次更新视图集序列,必先生成 $\mathcal{V}$ 中所有视图的依赖层次 $\mathcal{V} = \{\mathcal{V}^1, \mathcal{V}^2, \dots, \mathcal{V}^K\}$,使依赖传递地发生在不同层次之间。我们基于粗集理论^[11]对导出查询的视图集 $\mathcal{Q}$ 和中间视图集 $\mathcal{V}$ 建模,并使用约简动态生成 $\mathcal{V}$ 。

3.1 粗集约简

粗集理论为数据挖掘中对象区分和推理提供了一个理论框架,其核心是数据约简。对于由上述非空对象集 $\mathcal{Q}$ 和属性集 $\mathcal{V}$ 构成的信息系统 $\mathcal{A}(\mathcal{Q}, \mathcal{V})$,可以建立一个视图依赖的映射函数 $V_k(Q_i) \rightarrow [0, 1]$ 。其中, $Q_i \in \mathcal{Q} | i < n$ 并且 $V_k \in \mathcal{V} | k < m$; $V_k(Q_i)$ 在 $[0, 1]$ 取值,当 $Q_i \leq V_k$ 时, $V_k(Q_i) = 1$;否则 $V_k(Q_i) = 0$ 。如表1左侧所示:任意对象 $Q_i, Q_j \in \mathcal{Q}$ 在 $\mathcal{V}$ 下是不可分辨的,当且仅当 $\forall V_k \in \mathcal{V}, V_k(Q_i) = V_k(Q_j)$,称由其决定的关系 $IND(\mathcal{V})$ 为不可分辨关系。 $IND(\mathcal{V})$ 具备自反、对称和传递性,所以 $\mathcal{V}$ 是一族等价关系。 $\mathcal{Q}/(IND(\mathcal{V}))$ 为 $\mathcal{Q}$ 在 $\mathcal{V}$ 下的所有等价类,辨识矩阵 $M_D(\mathcal{V})$ 为一个 $n \times n$ 阶矩阵,每个元素均为属性集的析取 $\vee m_D(i, j)$,其中 $m_D(i, j) = \{V \in \mathcal{V} | \forall i, j=1, \dots, n (V(Q_i) = V(Q_j))\}$ 。

替代或消除任意查询谓词均不会引起查询执行代价增加。

由式(6)提供的单调代价模型, $V = \{v^1, v^2, \dots, v^K\}$ 可满足:

性质 2(层间依赖性)

$\forall (v^I, v^J \in V) \wedge (V_i \in v^I) \wedge (V_j \in v^J), I > J \rightarrow V_j \not\in V_i$;

如 3.2 节的中间视图集构造可知,待选集 V 中包含所有查询的共享子集及其补子集。由式(6)的单调性可知:由式(5)选取的约简层内视图比同层中其他约简包含更少连接关系,故不依赖后续视图集。式(4)可保证:对于 $\forall (v^I \in V) \wedge (V_i \in v^I)$, $TotalCost(\mathcal{Q} \mid \bigcup_{j \leq I} v^j) < TotalCost(\mathcal{Q} \mid \bigcup_{j \leq I} v^j - \{V_i\})$, 使得各层仅保留能够提高系统整体更新性能的视图。

4 算法实现

对于查询视图维护的中间结果共享,粗集约简有利于发现层内独立、层间依赖的中间视图层的集合。利用视图层的有效代价,我们论证了中间视图的各层之间的有序依赖,进而利用层内视图的联合代价估算获取整体更新性能最优的层次视图集 V 。因为基于单调代价模型假设,我们对 *MiniCost* 的计算依据 V 中视图层,由高到低搜索给定视图 V_i 或 Q_i 的可用共享视图,以避免完全匹配^[4]引起的估算代价过大。算法大体上分为候选视图生成、约简层选取、层内过滤三步:

显。如何对待选中间视图预先进行过滤以缩减考察空间,将作为我们进一步的研究方向。此外,算法对非均衡数据源的更新处理也有待进一步研究。

参 考 文 献

- Gupta A, Mumick I. Maintenance of materialized views: problems, techniques, and applications. *IEEE Data Eng Bull*, 1995, 18(2): 3~19
- 陈长清,程恩,冯玉才.一种用于搜索可能响应查询的候选实化视图的索引. *计算机科学*, 2003, 30(5): 176~179
- Salem K, Beyer K S, et al. How to roll a join: asynchronous incremental view maintenance. In: *Proceedings of SIGMOD*, 2000, 129~140
- Lee K Y, Son J H, et al. Efficient incremental view maintenance in data warehouse. In: *Proceeding of CIKM*, 2001, 249~356
- Liu B, Rundensteiner E A, Finkel D. Maintaining large update batches by restructuring and grouping. *Information Systems*,

2007 (Accepted)

- Lee K Y, Kim M H. Optimizing the incremental maintenance of multiple join views. In: *the Proceeding of DOLAP*, 2005, 107~113
- 王新军,洪晓光,等.数据仓库中多数据源物化视图的一种有效更新算法. *计算机研究与发展*, 2004, 41(5): 874~879
- Labio W J, Yerneni R, et al. Shrinking the warehouse update window. In: *Proceedings of SIGMOD*, June 1999, 383~395
- DeHaan D, Larson P-A, et al. Stacked indexed views in Microsoft SQL server. In: *Proc of SIGMOD*, 2005, 179~190
- Folkert N, Gupta A, et al. Optimizing refresh of a set of materialized views. In: *the Proceeding of 31st VLDB*, 2005, 1043~1054
- Skowron A, Rauszer C. The discernibility matrices and function in information system. In: *Slowinski R, ed. Intelligent Decision Support Handbook of Application and Advances*