

数据仓库中物化视图选择策略

林小静, 薛永生

(厦门大学 计算机系, 福建 厦门 361005)

摘要: 为了提高决策支持和 OLAP 查询的响应效率, 数据仓库多采用物化视图的思想。因此, 物化视图的选择策略是数据仓库研究的重要问题之一。其目标是选出一组存储、维护代价与查询代价的总和为最小的物化视图。提出一个以 MVPP(multi-view processing plan)为视图选择的搜索空间的物化视图选择新算法——VSMF(views selection base on multi-factor)算法。该算法在存储空间约束下同时实现多查询最优化和视图维护最优化。

关键词: 数据仓库; 物化视图; 选择策略; 维护策略; 存储空间约束

中图法分类号: TP311.131 **文献标识码:** A **文章编号:** 1000-7024 (2007) 13-3056-04

Selection strategy of materialized views in data warehouse

LIN Xiao-jing, XUE Yong-sheng

(Department of Computer Science, Xiamen University, Xiamen 361005, China)

Abstract: A set of materialized views are stored in the data warehouse for the purpose of efficiently implementing decision-support or OLAP queries. The selection of materialized views is one of the most important issues in the data warehouse development. The goal is to select an appropriate set of views so that the total cost of storage, maintenance and query is minimized. A new algorithm named VSMF (views selection base on multi-factor) algorithm using multi-view processing plan structure as search space is proposed, which solve the problem considering both multi-query optimization and the maintenance process optimization under the storage space constrain.

Key words: data warehouse; materialized view; selection strategy; maintenance strategy; storage space constrain

0 引言

当前, 数据仓库领域的一个研究热点就是物化视图的选择问题。物化视图是指将查询视图预先计算并以表的形式存储在数据仓库中, 当执行 OLAP 查询时, 可直接从物化视图中获取查询结果, 避免了对底层数据作复杂的综合操作, 从而有效提高查询响应速度。因此, 物化视图是提高系统多维分析性能的有效手段。

但是, 物化视图也带来了大量存储空间和视图维护的开销, 必须在缩短响应时间和资源限制二者之间权衡。由此, 物化视图选择的目标就是, 在空间限制下, 选出一组恰当的视图物化, 使得其对一组查询的总查询代价和其自身的维护代价之和为最小。该问题为组合优化问题, 其已经被证明为是 NP-完全问题。这个问题的最优解的复杂度是 $O(2^n)$, 其中, n 是数据仓库中视图的总数。目前存在许多算法, 基于各自不同的代价计算模型, 通过各种途径求解该问题的近似最优解。

1 相关工作

文献[1]以视图大小为选择准则提出 PBS 算法, 在特定前

提下, 使复杂度降低为 $O(n \log n)$; 文献[2]和文献[3]将遗传算法的获取最优解的能力应用于最优物化视图集的选取, 并在降低算法复杂度方面进行了研究; 文献[4]提出基于单位空间上查询频率的 FPUS。

以上算法都仅仅从视图所占空间大小这一限制条件着眼, 而忽略了物化视图的维护时间。而在实际应用中, 随着数据存储技术的飞速发展, 视图的维护时间逐渐成为限制数据仓库不能物化所有视图的主要因素。

文献[5]提出了以物化视图总维护时间为约束条件的贪心算法 ITGA; 文献[6]综合考虑了查询代价和维护代价, 并提出了 MVPP(multi-view processing plan)作为视图选择的搜索空间以获得最优解; 文献[7]在 MVPP 的基础上, 应用遗传算法求解; 文献[8]提出了一种结合遗传算法和模拟退火算法的混合算法; 文献[9]结合与或图, 贪心算法以及 A*启发式算法探讨该问题的求解方法。以上算法, 都同时考虑了查询代价和维护代价, 但都集中在获取最小代价的多查询最优化, 而忽略了物化视图维护策略的优化对这个问题的影响。

文献[10]提出的代价模型考虑了使用不同视图维护策略而产生的最小维护代价, 然而, 在该代价模型中, 没有考虑这

收稿日期: 2006-07-20 E-mail: lovebysea@126.com

基金项目: 福建省自然科学基金项目 (A0310008); 福建省重点科技基金项目 (2003H043)。

作者简介: 林小静 (1982-), 女, 福建连江人, 硕士研究生, 研究方向为数据仓库、数据挖掘及分布式数据库等; 薛永生, 男, 教授, 研究方向为数据库理论与应用、分布式数据库、数据仓库、数据挖掘、网络技术等。

些视图的查询代价。

因此,Naha A.R.Yousri 和 Khalil M.Ahmed 同时考虑多查询优化和视图维护优化这两个问题,提出了 IRVSA 算法和 IMDVSA 算法^[10],但这两个算法都忽略存储空间约束。此外,IMDVSA 算法只考虑,使用增量更新策略的情况;而 IRVSA 算法虽然同时考虑两种维护策略,但其与 BPUS 算法主要缺陷相同:首先,其每一步选择都要重新计算所有待选择视图的增益;其次,其每一步选择的视图都将作为将来要物化的视图,没有考虑每选择出一个新的视图后,已选视图的增益将出现衰减,而这一变化可能会导致其应该从已选视图中删除。

综上所述,目前已有的算法大多只考虑一个条件约束,或者只考虑存储空间约束,或者只考虑视图维护时间约束,而后者又大多忽略了视图维护策略的优化所带来的影响。在存储空间和视图维护时间共同制约下,同时考虑查询代价、视图维护代价及视图维护策略优化对维护代价的影响的物化视图选择问题,目前没有提出相应的算法。

本文在上述问题上进行了更为深入的探索,基于 SPJ(select-project-join)视图假设的关系数据库模型,以 MVPP 为搜索空间,综合考虑存储空间、视图维护开销、视图维护策略优化及查询性能,提出了 IMDVSA 的改造算法——VSMF 算法。

2 问题描述

基于 SPJ 视图假设的数据仓库中,在存储空间 Space 的限制下,从 MVPP 中选择一个视图集合 M 加以物化,使得查询集合 Q 的查询代价和物化视图集 M 的维护代价之和最小。在此,我们假设数据仓库会周期性地更新,并且增量更新和重新计算两种策略在该数据仓库中同时使用。

3 物化视图选择算法—VSMF 算法

3.1 相关定义

定义 1 MVPP(multi-view processing plan): MVPP 是通过结合给定查询集 Q 中每个查询的最优方案构建起来的,它以有向无环图的形式来描述针对查询集 Q 的一个查询处理策略。

如图 1 所示, MVPP 的叶子结点相当于数据仓库中的基表,其根结点相当于一个查询的最终结果,其所有中间结点及根结点都定义为一个视图。以下的讨论都将基于 MVPP。

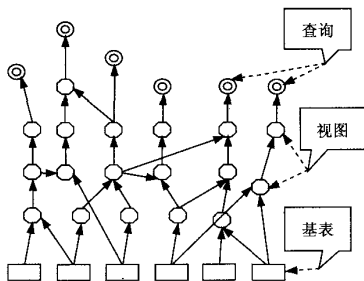


图 1 MVPP

定义 2 子孙结点:视图 u 为视图 v 的子孙结点,当且仅当在 MVPP 中,从视图 u 有一条通路到达视图 v。特别的,若通路中只有一条边,则称 u 为 v 的儿子结点, v 为 u 的父亲结点。

定义 3 一个查询 q 的查询代价 C(q,M): 当数据仓库中物化视图集为 M 时, q 的查询代价, 等于查询 q 对 M 中所有视图的查询代价的最小值。即: $C(q,M)=\min\{c(q,v), v \in M\}$ 。

定义 4 查询集 Q 的总查询代价 TQC(Q,M): 当数据仓库中的物化视图集为 M 时, Q 中所有查询的查询代价和该查询提交频率的乘积的总和。即: $TQC(Q,M)=\sum_{q \in Q}(f_q(q)*C(q,M))$ 。这里, $f_q(q)$ 是查询 q 的提交频率。

定义 5 物化视图 v 的维护代价 MC(v,M): 基于数据仓库中同时使用增量更新和重新计算两种维护策略, 因此 v 的维护代价为当数据仓库中物化视图集为 M 时, 分别使用两种策略所需代价的最小值。即: $MC(v,M)=\min\{IMC(v,M), RMC(v,M)\}$ 。

其中, IMC(v,M) 为使用增量更新策略所需的维护代价, RMC(v,M) 为使用重新计算策略所需的维护代价。二者均可由文献[10]的计算公式求得。

此处特别提到以数据仓库中物化视图集 M 为前提, 这是因为, 当使用增量更新策略时, 物化视图集 M 是否包含视图 v 的子孙结点, 与视图 v 的维护代价大小关系密切, 这将后文中详细说明。

定义 6 物化视图集 M 的总维护代价 TMC(M): M 的总维护代价为 M 中每个视图 v 的维护代价与该视图更新频率的乘积的总和。即: $TMC(M)=\sum_{v \in M}(f_v(v)*MC(v,M))$ 。其中, $f_v(v)$ 是视图 v 的更新频率。

定义 7 物化视图集 M 的总代价 TC(M): M 的总代价为当 M 被物化时, M 的总维护代价和查询集 Q 的总查询代价的加权和。即: $TC(M)=\delta*TMC(M)+TQC(Q,M)$ 。

将 TMC(M) 乘以权重因子 $\delta(\delta>0)$, 是因为不同的系统对查询性能和视图维护性能的要求不同, 系统管理员可以自行设定权重。

定义 8 视图 v 的增益 B(v,M): 视图 v 的增益为当物化视图集为 M 时的总代价与物化视图集为 $M \cup \{v\}$ 时的总代价的差值。即: $B(v,M)=TC(M)-TC(M \cup \{v\})$ 。

定义 9 视图 v 的大小 S(v): 视图 v 的大小即其物化所需的存储空间, 可采用文献[12]提出的基于数学估算和无重复采样的 Sample Frequency 算法估算。

定义 10 视图 v 单位空间的效益 BS(v,M): 物化视图集为 M 时, 视图 v 的增益与其所占存储空间的商。即: $BS(v,M)=B(v,M)/S(v)$ 。

3.2 算法理论基础

当采用增量更新策略进行物化视图的维护时, 物化一些额外的视图将降低维护视图集的代价^[10]。当更新视图集时, 一个物化视图的所有子孙结点由于基表的改变而产生的更新将传播到该物化视图。儿子结点的更新被用于计算父亲结点的更新。在某些情况下, 例如: 连接操作, 不单是儿子结点的更新数据参与计算, 而是儿子结点的全部数据都需要参与计算。于是, 如果儿子结点已经被物化, 则它的计算代价就节省了。同时, 由于儿子结点的更新要被用来计算父亲结点的更新, 因此物化儿子结点而产生的维护代价就成为父亲结点维护代价的一部分, 即不论儿子结点是否物化该结点, 都会产生这部分代价。

IMDVSA 算法正是基于这个理论提出的。该算法的优点

是复杂度低, 仅为 $O(n)$ 。但是, 该算法仅仅考虑增量更新策略, 一旦一个视图确定将被物化, 其所有的子孙结点都将被物化。其忽略了重新计算策略更优的可能性, 若一个视图重新计算的维护代价更低, 则物化该视图所有子孙结点只会增加系统的维护开销。

另外, 该算法还忽略了存储空间约束。在极端情况下, 该算法所选出的物化视图集 M 可能包含几乎所有的候选视图, 从而导致庞大的存储空间和视图维护代价。

基于上述考虑, 提出了 IMDVSA 算法的改进算法——VSMF (views selection base on multi-factor) 算法及考虑存储空间约束对物化视图进行调整的算法——MVSCA (modulation of views under space constraint algorithm) 算法。

3.3 物化视图的选择

对于一组查询集合 Q , 首先使用文献[6]中提到的算法构造出 MVPP, 接着使用 VSMF 算法, 从 MVPP 中求得视图集后, 运行 MVSCA 算法调整物化视图集, 使之满足存储空间约束。

算法中 M 表示选出的物化视图集, N 表示未搜索的物化视图集, n 表示 MVPP 中的层次, C 表示候选视图集, $D(v)$ 表示视图 v 的所有子孙结点, TS 表示物化视图集 M 所需的存储空间。

算法 1: VSMF 算法

输入: MVPP, 查询集 Q

输出: 应被物化的视图集 M

过程: 按广度优先策略搜索 MVPP 中的视图, 计算访问的当前结点(视图)的增益(按定义 8 的公式计算), 增益大于 0, 则物化该视图, 同时判断该视图采用哪种维护策略更优, 若使用增量更新策略更优, 则物化该视图的所有子视图, 接着将访问过的结点及要物化的结点从搜索空间中删除。

VSMF(MVPP, Q)

Begin

$M = \{\text{MVPP 中所有的视图}\};$

$N = \{\text{MVPP 中所有的视图}\};$

$C = \{\}; n = 0;$

Repeat

$C = \{\text{所有第 } n \text{ 层的视图}\} \cap N;$ // 未搜索的视图中最上层的视图集

对 C 中的每个视图 v

If $B(v, M) < 0$

Then {

$M = M - \{v\};$

$N = N - \{v\};$ }

Else If $IMC(v, M) < RMC(v, M)$

Then

$N = N - \{v\} \cup D(v);$

Else $N = N - \{v\};$

$n = n + 1;$

Until $n = \text{MVPP 的高};$

Return MVSCA(Space, M);

End

算法 2: MVSCA 算法

输入: 物化视图集 M , 存储空间限制 S ;

输出: 满足空间约束的物化视图集 M ;

过程: 估算物化视图集 M 需要的存储空间, 若超过空间限制, 则采用贪心算法对 M 进行调整。首先计算每个视图的单位空间的增益。一个视图增益越小, 所占空间越大, 则其单位空间的增益越小, 因此, 每次选择单位空间的增益最小的视图从 M 中删去, 直到 M 满足空间约束。

MVSCA(Space, M)

Begin

$TS = \sum S(v), v \in M;$

If $TS < \text{Space}$

Then Return M ;

Else

Repeat

计算 M 中每个视图 v 单位空间增益 $BS(v, M)$;

$v = M$ 中 $BS(v, M)$ 最小的视图;

$M = M - \{v\};$

$TS = TS - S(v);$

Until $TS < \text{Space};$

Return M ;

End

3.4 算法分析

VSMF 算法综合考虑了查询性能、存储空间约束、物化视图维护时间约束及不同更新策略的影响等因素。算法使用的增益计算模型较其它算法更为合理。算法采用从视图集 M 中逐个删去增益为负的视图的方法, 从视图维护代价方面分析, M 中保留的视图或者是采用重新计算策略, 或者是采用增量更新策略, 后者的所有子孙结点皆保留物化视图中, 因此, 删去的视图不会影响保留下视图的维护代价; 从查询性能方面分析, 物化视图集中视图总数减少, 保留下的每个视图对系统查询性能贡献将提高或者不变。

由此可知, M 中保留下的视图的增益不会随着 M 中视图数量的减少而降低, 从而避免了 BPUS 等算法中出现的视图增益衰减的弊病。算法同时考虑增量更新和重新计算两种维护策略, 对使用重新计算策略更优的视图, 避免了物化其对提高查询性能没有贡献的子孙视图, 节省了大量视图维护开销, 同时也降低了存储空间代价。最后, VSMF 算法调用 MVSCA 算法, 调整物化视图集, 使之满足空间限制 Space。VSMF 算法的复杂度为 $O(n) + \text{MVSCA 算法的复杂度}$ 。MVSCA 算法的复杂度主要来自于选择最小单位空间增益视图的操作, 选择恰当的算法, 复杂度为 $O(n \log n)$ 。因此, VSMF 算法的复杂度最差情况下为 $O(n \log n)$, 最优情况下, 即当给定的空间约束足够大而无需调整视图集, VSMF 算法的复杂度为 $O(n)$ 。

4 实验及比较

4.1 实验设计

目前各种物化视图选择算法中, 使用 MVPP 为搜索空间且同时考虑查询代价和维护代价的算法主要有 YKL 算法^[6]、IMDVSA 算法^[1]、IRVSA 算法^[11]等, 为了实验结果更具代表性, 选择考虑不同更新策略的 IRVSA 算法及只考虑增量更新策略的 IMDVSA 算法与 VSMF 算法做比较。

测试环境：硬件平台：P4 3.0GHz, 1G RAM; 操作系统：Windows 2003 Server; 数据库平台：Microsoft SQL Server 2000; 算法使用 Jbuilder2006 实现。

4.2 实验分析与对比

我们所选择的测试数据集包含一个事实表，每个维表都有 3 个层次。我们每次都利用模拟查询发生器产生 2 000 次查询时间，查询的分布满足 2~8 原则，即 80% 的查询量产生于 20% 的查询。对基表更新频率的设定，采用赋予 0~1 之间的随机数的方法。我们从算法时间开销，结果集 M 对查询的平均响应速度，及维护时间 3 个方面进行比较分析。从算法时间开销比较，实验结果如图 2 所示。

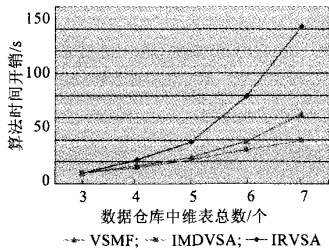


图 2 算法时间开销比较

可以看出，VSMF 算法相对于 IRVSA 算法，算法时间开销很低，相对于 IMDVSA 算法，随着维表的数目增长，VSMF 算法时间开销增长要快些，这是由于当选出的视图集不满足空间约束时，算法会对视图集进行调整。因此，如果给定的空间约束足够大，VSMF 算法的时间开销将不会比 IMDVSA 算法大。结果集对查询的平均响应速度比较，实验结果如图 3 所示。可以看出，3 种算法得到的物化视图集的平均查询响应速度相差不多。

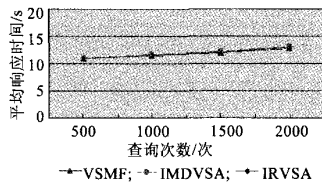


图 3 平均响应时间比较

维护时间的比较：首先从数据仓库的基表中随机选取 10% 作为更新基表，之后每次更新基表的数量以 10% 递增。设定每张表更新数据量为该表数据量的 10%，可得到实验结果如图 4 所示，可以看出，VSMF 算法的视图维护性能优于 IM-

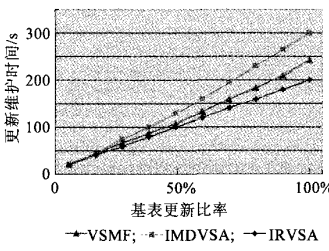


图 4 更新维护时间比较

DVSA 算法，而较 IRVSA 算法略差，但其在算法时间开销上的优势足以弥补这一缺陷。

5 结束语

本文提出的 VSMF 算法综合考虑数据仓库中影响系统查询性能和维护性能的各种因素，使用更为合理的增益估算模型和广度优先搜索，并使用 MVPP 削减算法的搜索空间。实验和分析表明算法的复杂度低而得到结果集具有较好的查询响应性能和较低维护开销。

参考文献：

[1] Clarke I, Sandberg O, Wiley B, et al. Freenet: A distributed anonymous information storage and retrieval system[C]. Proc of the Workshop on Design Issues in Anonymity and Unobservability. Berlin: Springer-Verlag, 2001: 46-66.

[2] Joseph S R H. NeuroGrid: Semantically routing queries in peer-to-peer networks[C].Pisa: International Workshop on Peer-to-Peer Computing, 2002:78-90.

[3] Tang Chunqiang, Xu zhichen, Dwarkada S. Peer-to-peer information retrieval using self-organizing semantic overlay networks [C]. Karlsruhe, Germany: Proc of SIGCOMM Conf, 2003.

[4] Cohen E, Fiat A, Kaplan H. Associative search in peer to peer networks: harnessing latent semantics [C]. The 22nd Annual Joint Conf of the IEEE Computer and Communications Societies. California: IEEE Computer Society Press, 2003: 1261-1271.

[5] Sripanidkulchai K, Maggs B, Zhang H. Efficient content location using interest-based locality in peer-to-peer systems [C]. Proc of Infocom, 2003.

[6] Yang J, Karlapalem K, Li Q. Algorithms for materialized view design in data warehousing environment[C]. Athens, Greece: Proc of the 23rd International Conference of Very Large Data Bases,1997:136-145.

[7] Horng Jorng-Tzong, Chang Yu-Jan, Lin Baw-Jhiune, et al. Materialized view selection using genetic algorithms in a data warehouse system[C]. Washington: Proc of the Congress of Evolutionary Computation, 1999: 2221-2227.

[8] 徐海涛,郑宁.数据仓库中物化视图选择的一种混合算法[J].计算机工程与设计,2005,26(10):194-197.

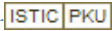
[9] Himanshu Gupta, Inderpal Singh Mumick. Selection of views to materialize in a data warehouse[J]. IEEE, 2005(17):24-43.

[10] Mistry H, Roy P, Sudarshan S, et al. Materialized view selection and maintenance using multi-query optimization [C]. Proceedings of SIGMOD'01, 2001: 307-318.

[11] Noha A R Yousri, Khalil M Ahmed, Nagwa M Ei-Makky. Algorithms for selecting materialized views in a data warehouse[J]. IEEE, 2005(5):27-35.

[12] Runapongsa K, Nadeau T P, Teorey T J. Storage estimation for multidimensional aggregates in OLAP [C]. Proc of the 10th CASCON Conf. Orlando, USA: IBM Press, 1999:40-54.

数据仓库中物化视图选择策略

作者: 林小静, 薛永生, LIN Xiao-jing, XUE Yong-sheng
作者单位: 厦门大学, 计算机系, 福建, 厦门, 361005
刊名: 计算机工程与设计 
英文刊名: COMPUTER ENGINEERING AND DESIGN
年, 卷(期): 2007, 28(13)
被引用次数: 1次

参考文献(12条)

1. Clarke I, Sandberg O, Wiley B Freenet: A distributed anonymous information storage and retrieval system 2001
2. Joseph S R H NeuroGrid: Semantically routing queries in peerto-peer networks 2002
3. Tang Chunqiang, Xu zhichen, Dwarkada S Peer-to-peer information retrieval using self-organizing semantic overlay networks 2003
4. Cohen E, Fiat A, Kaplan H Associative search in peer to peer networks: harnessing latent semantics 2003
5. Sripanidkulchai K, Maggs B, Zhang H Efficient content location using interest-based locality in peer-to-peer systems 2003
6. Yang J, Karlapalem K, Li Q Algorithms for materialized view design in data warehousing environment 1997
7. Horng Jorng-Tzong, Chang Yu-Jan, Lin Baw-Jhiune Materialized view selection using genetic algorithms in a data warehouse system 1999
8. 徐海涛, 郑宁 数据仓库中物化视图选择的一种混合算法[期刊论文]-计算机工程与设计 2005(10)
9. Himanshu Gupta, Inderpal Singh Mumick Selection of views to materialize in a data warehouse 2005(17)
10. Mistry H, Roy P, Sudarshan S Materialized view selection and maintenance using multi-query optimization 2001
11. Noha A R Yousri, Khalil M Ahmed, Nagwa M Ei-Makky Algorithms for selecting materialized views in a data warehouse 2005(05)
12. Runapongsa K, Nadeau T P, Teorey T J Storage estimation for multidimensional aggregates in OLAP 1999

相似文献(10条)

1. 期刊论文 衣振萍, 潘昌景, 郭强, 姜斌, YI Zhen-ping, PAN Jing-chang, GUO Qiang, JIANG Bin 数据仓库环境中可扩展的动态物化视图选择方法 - 计算机与现代化 2007, "" (8)

数据仓库通常要对大量的数据进行运算, 以精简的结果来回答用户的查询, 这一特点使得物化视图技术在数据仓库中尤为重要。然而现有支持物化视图自动选择的方法是静态的, 它违背了联机分析处理(OLAP)和决策支持系统(DSS)的动态本质。本文提出了可扩展的动态物化视图方法, 通过将整个物化视图选择问题(MVS)分解为三个阶段, 降低了问题的复杂度, 提高了物化视图的有效性。通过动态调整, 物化视图能即时适应查询需求。算法复杂度分析证明了方案的可扩展性。动态调整算法模拟实验验证了方案具有很好的自适应性。

2. 学位论文 衣振萍 数据仓库中基于访问频率的动态物化视图的研究 2005

数据仓库的在线分析处理(OLAP, On-Line Analytical Processing)和在线数据挖掘(OLDM, On-Line Analytical Mining)通常要对大量的数据进行运算, 以精简的结果来回答用户的查询。数据仓库系统的这一特点使得物化视图技术在数据仓库中尤为重要。物化视图是经过数据预处理而生成的表, 这些表物理地存储在数据仓库中, 通过对其简单运算或简单查找回答用户查询, 从而能够在很大程度上提高数据仓库的查询响应速度。由于物化视图占用存储空间、需要进行更新维护, 所以物化所有的查询对应的视图是不现实的, 必须考虑选出哪些视图进行物化, 这就是物化视图选择问题(MVS, Materialized Views Selection)。

尽管对于MVS问题已经有大量的研究, 然而现有研究还达不到工业中所要求的稳定性、健壮性, 数据仓库的商业产品对物化视图自动选择支持不够理

想。现有支持物化视图自动选择的数据仓库产品采用的是静态物化视图选择方案，这种方案违背了OLAP和决策支持系统(DSS, DecisionSupportSystem)的动态本质。而数据仓库未来的发展对物化视图选择的效率、易用性、有效性和自适应性提出了更高的要求。

本文提出了一种基于视图访问频率的动态物化视图方案，该方案能够克服静态物化视图选择方案的缺点，它具有自适应、高效、易操作的特点，能解决大规模的MVS问题。

方案根据视图不同的访问需求特征，在不同阶段、以不同的方式物化视图，从而降低了整个MVS问题的复杂度，提高了给定存储空间中的物化视图的有效性。系统调用多项式时间的改进的贪心算法，自动选出初始物化视图，填充部分物化视图存储空间，通过对MVS本阶段的子问题规模进行控制，提高本阶段的执行效率。方案认为视图的访问频率反映了用户的查询趋势，因此构造了以视图访问频率为主要因素的收益模型，并以此模型计算的收益值作为物化视图的调整标准，对物化视图集动态物化和调整。这样物化视图集能够随着用户查询趋势的改变而改变，具有自适应性。方案把物化视图分为两种：永久物化视图和临时物化视图，只有临时物化视图才可以被调整，从而避免了具有较高稳定访问频率的物化视图被删除。物化视图预警线(PWL, PriorWarningLine)的引入，可以提前发现物化视图存储空间即将被填满的状态，提前进行物化视图调整前的准备，从而提高系统响应查询的效率。本文提出了动态物化选择及调整的核心算法，采用TPC-H基准数据模式、用1GB的数据构造和填充了Oracle数据库，以此为基础进行实验查询和分析对比，实验验证了收益模型和动态物化调整算法的有效性。

本文提出了动态物化视图方案具有自适应、高效、易操作的特点，但要将其完全应用于数据库库中，还需要考虑与其它机制的协调，以及一些集成的细节，这也是下一步要研究的内容。

3. 期刊论文 [姜合, 杨春花, 耿玉水, Jiang He, Yang Chunhua, Geng Yushui 超市数据仓库中物化视图的选择与调整策略 - 计算机应用与软件 2007, 24 \(3\)](#)

物化视图选择是数据库研究领域的一个重要课题,其选择策略直接影响到数据库的查询效率.通过对超市数据库的设计及已有研究成果的分析,对物化视图的选择算法做了一些改进,并给出了一种据查询情况的变化动态调整物化视图集的算法.

4. 学位论文 [秦智平 数据库结构设计与物化视图选择 2003](#)

该文在对现有数据库的概念、特征、数据组织结构和设计方法进行分析的基础上,提出了一种数据库体系结构,将数据库分析环境中的数据划分为核心数据、全局扩展数据、局部扩展数据和私有扩展数据四类,并分别对这四类数据的结构设计问题进行了研究,将物化视图选择理论用于后三类扩展数据的结构设计中.该文针对现有物化视图选择理论在数据库工程实践中面临的困难,提出了查询泛化的概念,并给出了两种查询泛化的算法;在此基础上,针对各类扩展数据的特点,恰当地选择和改进物化视图选择问题上的现有理论成果,进行数据库的结构设计.数据库的结构应该与用户的分析需求相适应,而后者是复杂多变的,当用户的分析需求变化超出某一度时,需要对数据库的结构作恰当的调整,很明显的一种选择就是重新执行物化视图选择算法,形成新的物化视图集,但这带来的问题是无法控制调整量,而重建物化视图的开销极大,因此该文提出了在限定调整量的情况下,如何对物化视图集作局部调整的问题,并给出了相应的算法.最后,该文介绍了上述研究成果应用于上海一家大型超市数据库系统中数据集市部分的设计概况和部分细节,并对研究工作进行了总结和展望.

5. 期刊论文 [张晓辉, 袁愿, 虞健飞, 张恒喜 数据库物化视图选择的混合算法 - 计算机应用 2003, 23 \(7\)](#)

物化视图是提高数据库的查询响应能力以高效支持决策分析的重要手段,但物化视图集选择是一个复杂问题.结合启发式算法的快速收敛能力和遗传算法的全局优化能力的两层物化视图求解方案提供了物化视图选择问题求解的可行途径.

6. 学位论文 [姜全胜 数据库物化视图一致性维护研究 2008](#)

数据库是计算机信息化不断发展的产物,它将大量用于事务处理的数据库数据进行清理、抽取和转换,并按决策主题的需要重新进行组织,以达到快速有效支持决策的目标.物化视图的联机一致性维护技术是数据库联机维护技术研究中的一个热门问题.在数据库物化视图研究领域应用较多的是对物化视图的一致性维护问题,并且大部分研究都是基于视图定义在关系表主键的假设基础上而进行的,其中应用较为成熟的是ECA-Key算法和ECA-Key补偿算法,这两种算法在查询时直接利用源数据库关系,避免了查询时数据库与数据视图的不一致性.但是ECA-Key算法只在视图定义带源关键字并且更新查询的发出与接收的顺序保持一致时算法才成立,由于数据库物化视图应用的复杂化,网络环境下更新操作的频繁性和顺序上的不确定性,即由于业务分布、介质及网络通信等方面的原因,数据库收到的查询计算结果,和它向各数据源发出的计算查询顺序并不一定一致,从而引起更新维护后数据的不一致,这导致了ECA-Key算法和ECA-Key补偿算法应用面的狭窄,并逐渐显现出其弊端.并且算法ECA-Key采用对物化视图完全备份的方式进行,增加了维护与数据写回的开销.

本文分析了ECA-Key算法的应用示例,并在此基础上提出了关系数据表的扩展模式,在扩展模式上提出了物化视图更新算法Expansion(包括源数据库端的Expansion-DB算法和数据仓库端的Expansion-DW算法),算法的基本思想是将普通数据表定义进行模式上的扩展,用扩展字段记录更新操作进行的顺序与操作的类型,并根据一定的规则对源数据库的操作进行判断,然后对数据库物化视图端进行必要的修改操作,源数据库端和数据仓库端采用查询通知和反馈确认的方式进行,从而保证了视图维护事务的数据一致性.在介绍了算法思想之后给出了扩展模式下更新算法Expansion的应用示例,证明了算法的正确性.最后讨论了物化视图自维护方面的问题,给出了物化视图自维护的概念,讨论了物化视图自维护的特点,分析了物化视图自维护的判断依据,简单提出了物化视图自维护算法的基本思想,并给出了物化视图进行简单自维护的条件和物化视图向自维护方向的简单扩充.

7. 学位论文 [林小静 数据库中多维数据物化视图的选择 2007](#)

随着信息时代的来临,企业面临大量数据,如何快速从中提取信息、制定市场策略,以便对市场做出及时灵活的反应,成为企业在市场竞争中立于不败之地的关键.联机分析处理OLAP(Online Analytical Processing)正是用户获得决策支持的主要手段.

OLAP必须支持各种可能的查询,相当一部分查询可能要涉及大量的数据,并需要对数据进行选择、投影、连接等处理,这是一个非常耗时的过程,然而一个决策支持系统要求它的查询能够被快速响应.解决这一矛盾通常采用的一个有效的方法是:数据库针对OLAP可能的查询对原始数据进行选择、投影、连接等预处理,建立物化视图(Materialized View).但是,物化视图也带来了大量存储空间和视图维护的开销,必须在缩短响应时间和资源限制二者之间进行权衡,选择出恰当的物化视图集合.因此,物化视图的选择问题作为设计、构建数据库的关键问题之一,成为当前数据库领域的一个研究热点.此外,物化视图的相关研究还包括用物化视图改写查询、物化视图的维护以及物化视图的动态调整等.

本文主要针对基于关系数据库的OLAP系统中的多维数据物化视图的选择和动态调整问题进行研究,在提出一个更为合理的视图增益模型之后,分别提出基于MVPP的物化视图选择算法——VSMF算法,调整物化视图集使之满足空间约束的算法——MVSCA算法,物化视图实时调整算法——RMMV算法以及基于MVPP的物化视图动态调整算法——DMMF算法。

VSMF算法以MVPP为视图搜索空间,综合考虑了物化视图影响系统查询性能和维护性能的各种因素,使系统得到较好的查询性能和较低的维护开销.MVSCA算法根据视图的单位空间增益对物化视图集进行调整,使得其满足给定的空间约束.RMMV算法对物化视图集进行实时调整,避免了视图的重复计算和对视图大小的估算,提高了物化视图选择的效率和准确性,使系统在运行过程中能及时反映查询分布趋势、维持较好的查询响应性能.DMMF算法从MVPP的角度讨论物化视图集的动态调整,其综合了批量调整算法和实时调整算法的优点,同时避免了二者的缺陷.从实验结果和比较分析可以看出,以上算法具有一定的优越性。

8. 期刊论文 [杨少军, 范金存, 李庆忠 数据库中物化视图的选择 - 计算机应用 2003, 23 \(9\)](#)

物化视图是数据库中提高查询效率的有力方法,物化视图的选择一直是数据库领域的研究热点.通过对星型模型的研究,根据对数据库的常用查询及其执行概率,设计出一个候选视图的算法,并详细介绍了线性代价模型,在该模型和候选视图算法基础上,参照文献[4]提出一个改进的物化视图选择贪心算法.

9. 学位论文 [吕晓 基于聚类的动态物化视图选择研究 2009](#)

经过多年发展,数据库已广泛应用于各行业,随着时间的推移,数据库中的数据量迅猛增长,为了解决查询响应所需时间越来越长的问题,物化视图技术应运而生,并已成为数据库中的一个研究热点.物化视图技术将视图所对应数据加以实际物理存储,通过预计算的方式加快查询响应速度,然而,其本身也需要耗费大量的资源,因而如何选择一组合适的视图进行物化就成为数据库查询中的一个重要问题.现有的物化视图选择技术多为静态选择算法,在一定程度上与决策支持应用系统的动态特性相矛盾,而动态物化视图选择算法研究较少,且在系统开销过大的缺点.针对这两者的

不足,在前人研究的基础上,本文提出并实现了一个基于聚类的动态物化视图选择算法,该算法结合使用了所提出的静态物化视图改进算法与聚类改进算法。

本文在探讨了数据仓库、物化视图选择及聚类分析等技术的基础上,进行了基于聚类的动态物化视图选择方法研究,提出了一种基于聚类的动态物化视图选择算法CBD—MVS。该算法利用聚类技术来对数据仓库中的用户查询语句进行聚类,再对聚类后的各个簇中的用户查询语句进行合并,得到数量较少的候选物化视图,然后再选择一种合适的静态物化视图选择算法来得到最终的物化视图。

本文的主要研究内容为:

1. 针对现有聚类算法在对用户查询语句进行聚类处理的不足,把频繁闭项目集应用到聚类分析技术中,通过对用户查询语句执行频繁闭项目集挖掘算法,得到基于属性字段的关联规则,并根据这些规则求得属性字段的关联度矩阵和特征向量,计算出属性字段集相似度,执行k均值聚类算法获得聚类结果。实验表明该方法得到了较好的聚类结果。

2. 探讨了数据仓库技术及物化视图技术,着重研究了静态物化视图选择算法Greedy、BPUS和PBS,并分析其不足之处,提出了一种改进算法BGA。该算法使用启发式搜索算法的思想搜索格图,利用数据立方体格图之间存在的依赖关系,结合代价模型筛选出具有最大效益的物化视图,并将存储空间与新增效益共同作为阈值,在获得了与BPUS算法相同视图查询代价效果时,所耗费的时间明显少于后者。实验证明该算法是十分有效的。

3. 研究了数据仓库中物化视图的动态选择问题,针对现有物化视图选择算法的不足,提出了一种基于聚类的动态物化视图选择算法CBD—MVS。该算法采用基于频繁闭项目集的聚类算法对用户查询语句进行聚类,应用视图合并算法建立候选物化视图,利用改进的静态选择算法BGA生成最终应该被物化的视图。实验表明该算法是有效可行的。

10. 期刊论文 [朱文. 毛琴辉. 薛燕. 苏森. 张柏礼. ZHU Wen. MAO Qin-hui. XUE Yan. SU Sen. ZHANG Bai-li 数据仓库中物化视图维护算法的分析和比较—现代计算机（专业版）2008,“\(4\)](#)

随着数据源的更新,数据仓库中的物化视图必须得到及时的更新维护.而如何对物化视图进行高效的更新,以满足用户对查询响应速度和查询结果一致性、时新性的要求,这是数据仓库技术中非常复杂和重要的工作,也是一个迫切需要解决的关键性技术问题.以物化视图更新维护问题为主要研究对象,通过对现有各种维护算法深入的研究和分析,系统地进行了比较和总结,最后指出了该问题深入研究的方向.

引证文献(1条)

1. [张忠平. 张艳. 金晓丹. 何丽荣 一种基于中间结果集的有效视图维护算法\[期刊论文\]-计算机应用研究 2008 \(10\)](#)

本文链接: http://d.wanfangdata.com.cn/Periodical_jsjgcysj200713010.aspx

授权使用: 华南理工大学(hnlgdx), 授权号: 7009569c-1aa2-4d6b-93ac-9dfd018a1f24

下载时间: 2010年9月26日