

文章编号:1001-9081(2009)10-2772-02

基于时间段查询的物化视图策略

于 翔,印桂生

(哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001)

(yuxpointfly@tom.com)

摘要:物化视图是减少数据仓库中查询响应时间的有效方法。现有的物化视图选择策略主要考虑物化视图的初始选择方法以及动态更新方法。针对某时间段内查询进行物化视图更新的情况考虑不足,在贪心算法以及动态更新算法的基础上,提出了基于时间段内查询的物化视图更新策略。基于时间段查询的物化视图更新策略可充分适应用户需求,提高查询效率。

关键词:数据仓库;物化视图;查询;更新

中图分类号:TP311 **文献标志码:**A

Strategy of materializing views based on queries of time period

YU Xiang, YIN Gui-sheng

(College of Computer Science and Technology, Harbin Engineer University, Harbin Heilongjiang 150001, China)

Abstract: An effective method to reduce the response time to query is to materialize views in data warehouse. The current strategies mainly consider the methods of the primary selection of materialized views and the methods of dynamic updating. Concerning the situation that the updating of views based on queries of time period did not get enough attention, based on greedy algorithm and dynamic refreshment algorithm, the authors proposed a method of materializing views based on queries of time period. The method of materializing views based on queries of time period can meet the needs of clients, and enhance the efficiency of queries.

Key words: data warehouse; materialized view; query; refreshment

数据仓库是一个用来支持管理人员决策的面向主题的、集成的、非易失的且随时间变化的数据集合。为了以少量空间的代价换取数据查询响应时间的缩短,从而有效地支持对数据进行快速查询分析的需要,通常在数据仓库中进行视图的物化。而现有的物化视图选择策略主要考虑物化视图的初始选择方法以及如何动态地更新物化视图^[1-5],并未充分考虑用户对物化视图的具体应用。物化视图的选择问题是数据仓库设计中的研究热点,首先,要求数据仓库中的物化视图能够提供最佳的查询性能;其次,要求更新物化视图时的维护开销最小。如何权衡查询性能与维护开销,是物化视图选取时需要考虑的问题。本文在物化视图的动态选择算法^[6]的基础上,进一步考虑了有效满足用户的需要,针对具体时间段内查询的应用,提出了基于时间段查询的物化视图更新算法。

1 物化视图的方法选择

物化视图的选择主要有三种情况。

1) 物化所有可能存在的视图。这种方法可以缩短所有查询的响应时间,但此方法可能会浪费大量的存储空间,在数据库较大的情况下尤其明显。

2) 不进行视图的物化。这种方法不浪费额外的存储空间,但在响应查询时,直接从事实表中计算,在数据库较大的情况下,会耗费较长的时间去响应查询。

3) 物化一部分视图。通过考虑各种类型查询发生的概率不同、物化及更新各个视图所消耗的代价不同,对数据仓库中的一部分视图进行物化,既不浪费过多的存储空间,又能快速地响应大部分的查询请求,是一种比较灵活的折中算法。本文的物化视图更新策略是在第三种策略的基础上展开的。

由于实视图的目的是保存数据供查询访问,因而视图的表示与查询是一致的。常将查询和视图混用。

对于物化视图的初始选择我们采用贪婪算法^[1]来实现,经证明,贪婪算法所得到的实视图的效益不低于 $1 - 1/e$ 。

算法 1 贪婪算法(BPUS)

输入:可用空间 S

输出:物化视图的集合 M

WHILE $S > 0$ DO

FOR each v in V

计算 $B(v)$,并在其中找出值最大的 $B(v)$ 所对应的 v ;

IF $|v| \leq S$ THEN

$\{S = S - |v|; M = M \cup \{v\}; V = V - \{v\};\}$

ELSE $S = 0$;

ENDWHILE

RETURN M

其中: $|v|$ 表示视图 v 所占的空间; M 为物化视图的集合; S 为总空间的限制; V 为所有可能的数据节点,即视图的集合; $B(v)$ 表示对视图 v 进行物化,会获得的效益。

随着时间的变化,对物化视图的更新我们采用动态选择算法^[6],如下所示。

算法 2 物化视图的动态调整

输入:当前查询 q ,可用空间 S ,当前的物化视图集合 M ;
 $q \notin M$

输出:物化视图的集合 M

$V_{\text{remove}} = \emptyset$;

//将要删除的物化视图集合

$\text{search} = \text{TRUE}$;

//继续搜索标记

WHILE $S < q$ AND search DO

$v = M$ 中 $f(v)/|v|$ 最小的视图;

IF $f(v)/|v| < f(q)/|q|$ THEN

收稿日期:2009-04-28。 基金项目:国家 863 计划项目(2007AA01Z401);国家重点实验室基金项目。

作者简介:于翔(1978-),男,黑龙江哈尔滨人,博士,主要研究方向:数据仓库、数据挖掘;印桂生(1964-),男,江苏泰兴人,教授,博士生导师,主要研究方向:数据库和知识库应用系统、虚拟现实。

```

S = S + |v|;
V_remove = V_remove ∪ {v};
M = M - {v};
ELSE search = FALSE;
ENDWHILE
IF S ≥ |q| THEN
删除 V_remove 中的视图;
M = M ∪ {q};
物化 q; S = S - |q|; M = M ∪ {q};
ELSE M = M ∪ V_remove;

```

其中, $f(q)/|q|$ 表示单位空间的频率, $|v|$ 表示视图 v 所占的空间; M 为物化视图的集合; S 为总空间的限制; V 为所有可能的数据节点, 即视图的集合。

2 基于时间段查询的物化视图策略

例 1 考虑一个电信部门计费例子, 其中有三个维: 缴费时间、缴费细节、计费员工, 其度量量为费用 (Charge)。

```

Time(Time_ID, Day, Month, Year) //缴费时间
Fee(Fee_ID, Fee_Acc, Customer_Name, Fee_type) //缴费细节
Employee(Employee_ID, Ename, Edept) //计费员工
Charge(Time_ID, Fee_ID, Employee_ID, Charge_Amount) //事实表

```

维的等级情况如图 1 所示, 用多维数据格^[1]表示如图 2。

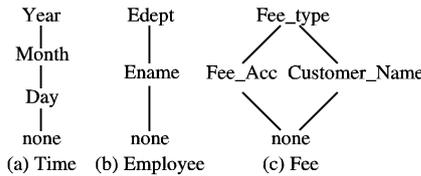


图 1 维的等级情况

图 2 所示的偏序图中, 每一个节点就代表一个视图。显然, 节点 (T, F) 可由节点 (T, F, E) 得到, 节点 (T) 可由节点 (T, F) 或 (T, E) 得到, 也可以由节点 (T, F, E) 得到。

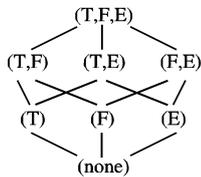


图 2 偏序图表示

经过某周期, 人们在不同时间段内的查询可能会具有某种相似性。因而, 在已有的物化视图集的基础上, 根据过去某段时间内查询的趋向而针对某种应用进行物化视图的替换是很有必要的。由于空间的限制, 需要淘汰已物化的视图集合中在一定时间内未被访问或访问较少的视图, 同时根据选定时间段内查询加入新物化的视图。类似于操作系统中的最近最久未访问的页面置换算法来淘汰已物化的视图, 进行面向应用的多维数据实视图的动态选择。鉴于以上情况, 我们提出了基于时间段查询的物化视图策略。

如在电信部门的计费过程中, 由于电信部门每月计算当月话费的时间具有一定的周期性和规律性, 基于时间段查询的物化视图策略可以充分根据以往月份的计费时间段内查询分布情况, 进行物化视图的替换, 以减少响应计费人员进行当月各项查询所需要的时间。

在以下定义的基础上, 提出基于时间段查询的物化视图策略。

定义 1 多维数据的查询集合: 在多维数据集合 MD 上的某段时间内的所有查询 $Q_{Interval}$ 。将 $Q_{Interval}$ 表示为 $\{q_1, q_2, \dots, q_n\}$, 其中 q_i 为第 i 个查询。

定义 2 物化视图集合的基视图: 在物化视图集合中, 基视图 v_β 是汇总级别最低的, 可利用它计算多维数据格中的任意节点。如图 2 中的 (T, F, E)。

定义 3 单位空间的访问频率^[6]: 某段时间内, 查询 q 的访问频率 $f(q)$ 与满足该查询的多维数据格节点 (即视图) 大小 $|v|$ 的比值 $f(q)/|v|$ 。

基于时间段查询的物化视图策算法描述如下:

算法 3 基于时间段查询的物化视图算法

输入: 特定周期内的所有查询 $Q_{Interval}$, 可利用空间的初始大小 S , 已物化视图的集合 M

输出: 新物化视图集合 M'

$M' = \Phi; M' = M' \cup v_\beta; Flag = 1;$ // 其中 v_β 为基视图

$S = S - |v_\beta|;$

FOR each $\{q\}$ in $Q_{Interval}$

计算出 $f(q)/|v|$ 并排序;

//按值由大到小排序

IF $S > 0$ AND $|q| < S$ THEN

$\{S = S - |q|; M' = M' \cup \{q\};\}$

ELSE RETURN M' ;

IF $S > 0$ THEN

FOR each $\{q\}$ in M

计算出 $f(q)/|v|$ 并排序;

//按值 $f(q)/|v|$ 由大到小的顺序开始

WHILE $S > 0$ AND $Flag! = 0$ DO

IF $q \notin M'$ AND $|q| < S$ THEN

$\{S = S - |q|; M' = M' \cup \{q\};\}$

ELSE $Flag = 0;$

ENDWHILE

RETURN M'

如果过去某段时间内的查询采用了动态物化视图策略, 若存在某一列查询与过去某特定时间段内的查询类似, 则可采用与过去特定时间段内的查询相同或相似的物化视图集合。即根据分布在过去不同时间段内的相似查询, 进行有针对性的视图物化, 以有效地响应当前查询。

3 策略应用举例

电信计费部门一般在每月月底进行当月费用结算工作, 在每年年底的某段时间进行年费用结算, 这个时间通常具有一定的周期性及规律性。根据动态更新策略对某月份费用结算期间的查询 (计费人员的查询), 进行视图的物化, 可以有效地减少查询响应时间。在下个月份的费用结算期, 可以根据上月份费用结算期间所淘汰及物化的视图来调整当前的物化视图集合, 以满足计费人员的查询。

考虑例 1 中电信部门计费的电信计费情况, 根据贪婪算法, 可以得到最初的物化视图集合。现假设各个候选视图的行数及大小分别为:

$Rows(T, F) = 5 \text{ MB}, Size(T, F) = 4 \text{ GB}; Rows(T, F, E) = 6 \text{ MB}, Size(T, F, E) = 6 \text{ GB};$
 $Rows(T, E) = 2 \text{ MB}, Size(T, E) = 2 \text{ GB}; Rows(F, E) = 4 \text{ MB}, Size(F, E) = 3 \text{ GB};$
 $Rows(T) = 1.5 \text{ MB}, Size(T) = 1 \text{ GB}; Rows(F) = 2.8 \text{ MB}, Size(F) = 2 \text{ GB};$
 $Rows(E) = 0.5 \text{ MB}, Size(E) = 0.5 \text{ GB};$
 总空间的限制设为 12 GB。

表 1 收益值的计算结果

候选视图	第 1 次/MB	第 2 次/MB	第 3 次/MB
(T, F)	3.0	2.0	1.0
(T, E)	12.0	—	—
(F, E)	6.0	4.0	—
(T)	4.5	0.5	0.5
(F)	3.2	3.2	1.2
(E)	5.5	1.5	1.5

利用贪婪算法进行对候选视图的选择, 收益值的计算结果如表 1 所示。
(下转第 2777 页)

明,本文给出的聚类算法能有效地进行用户浏览路径聚类。

当相似度阈值取为 0.8 时,得到 372 个类,删除只包含一条浏览路径的类,可得到有效的类为 13 个,如图 3 所示。

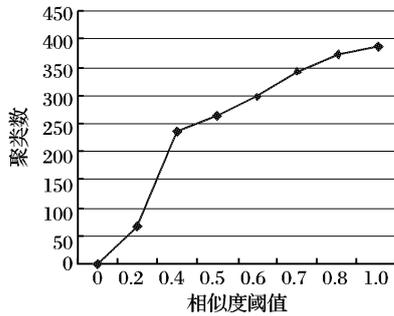


图 2 相似度阈值大小与聚类数变化关系



图 3 相似度阈值为 0.8 时典型匿名用户路径集

图 3 中 simpleSpecialUserPath 字段表示的是各个类的中心,是用 URL 的编号表示的典型匿名用户路径,SpecialUserPath 字段是以 URL 表示的典型匿名用户路径。

(上接第 2773 页)

算法的第一次循环节点 (T, E) 的收益计算过程为 $(6-2) \times 3 = 12$ MB,其余节点计算过程依次类推。经过三轮计算,选择每轮计算结果中收益最大的视图进行物化,并考虑所占空间的大小。第一轮计算选择收益最大的 (T, E) 视图进行物化,第二轮计算选择 (F, E) 视图进行物化,第三轮选择了 (E) 视图进行物化。此时,除去必须物化的基视图 (T, F, E) 所占用的 6 GB 空间。剩余部分所占用的空间总数为 2 GB + 3 GB + 0.5 GB = 5.5 GB。第四轮计算后选择的视图所占空间的大小为 2 GB,所以由于总空间的限制,不继续进行视图的物化。因此,经过三轮计算后,初始选择物化的视图为 (T, E), (F, E), 和 (E)。进而进行物化视图的动态调整。

若上个月份费用结算期间计费人员的所有查询 $Q_{Interval}$ 为 $\{q_1, q_2, \dots, q_n\}$, 现假设该周期内 (T, F)、(T, E)、(F, E)、(T)、(F)、(E) 的单位空间访问频率依次为 12%、15%、15%、17%、14%、2%, 根据基于时间段查询的物化视图算法可得到调整后的物化视图集合 (T, F, E)、(T)、(T, E)、(F, E)。则本月份可采用该物化视图集合来响应计费人员的查询。调整后的物化视图集合充分考虑了物化视图应用的实际情况,可以有效缩短查询响应时间,提高计费人员的工作效率。

4 结语

本文研究了在数据仓库环境下物化视图的选择问题,物化视图选择的好坏直接影响数据仓库的效率,在数据仓库数据模型的创建中具有重要意义。基于时间段查询的物化视图策略在贪婪算法和物化视图的动态选择算法^[6]的基础上,进一步考虑了查询集合的周期性和相似性。同贪婪算法相比较,它应用了视图的更新策略,可以更好地适应查询的变化,缩减查询的响应时间;同物化视图的动态选择算法相比,它考

虑不同时间段查询的情况,可以根据具体情况优化物化视图集合。

3 结语

Web 日志挖掘是一个较新的研究领域,具有广阔的发展和前景,文中所做的工作对于研究基于 Web 日志的匿名用户浏览路径挖掘技术以及构造实用的 Web 日志挖掘系统具有重要意义。

参考文献:

- [1] PALIOURS G, PAPTAEODOROU C, KARKALETSIS V, et al. Clustering the users of large Web sites into communities [C]// Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2000: 719 - 726.
- [2] FU Y, SANDHU K, SHIH M. Clustering of Web users based on access patterns [C]// International Workshop on Web Usage Analysis and User Profiling. San Diego: [s. n.], 1999: 142 - 162.
- [3] 姚洪波, 杨炳儒. Web 日志挖掘数据预处理过程技术研究[J]. 微计算机信息(管控一体化), 2006, 22(6-3): 234 - 236.
- [4] 宋擒豹, 沈钧毅. Web 页面和客户群体的模糊聚类算法[J]. 小型微型计算机系统, 2001, 22(2): 229 - 231.
- [5] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8: 338 - 353.
- [6] 业宁, 李威, 梁作鹏, 等. 一种 Web 用户行为聚类算法[J]. 小型微型计算机系统, 2004, 25(7): 1364 - 1367.

参考文献:

- [1] HARINARAYAN V, RAJARAMAN A, ULLMAN J D. Implementing data cubes efficiently[J]. ACM SIGMOD Record, 1996, 25(2): 205 - 216.
- [2] AGRAWAL R, GUPTA A, SARAWAGI S. Modeling multidimensional databases[C]// ICDE' 97: Proceedings of the 13 th International Conference on Data Engineering. Birmingham: IEEE Computer Society Press, 1997: 232 - 243.
- [3] BARALIS E, PARABOSCHI S, TENIENTE E. Materialized view selection in a multidimensional database[C]//VLDB' 97: Proceedings of the 23 rd International Conference on Very Large Data Bases. Athens: Morgan Kaufmann Publishers, 1997: 156 - 165.
- [4] SHUKLA A, DESHPANDE P, NAUGHTON J F. Materialized view selection for multidimensional datasets[C]//VLDB' 98: Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann Publishers, 1998: 488 - 499.
- [5] GUPTA H, HARINARAYAN V, RAJARAMAN A, et al. Index selection for OLAP[C]// ICDE' 97: Proceedings of the 13 th International Conference on Data Engineering. Birmingham: IEEE Computer Society Press, 1997: 208 - 219.
- [6] 谭红星, 周龙襄. 多维数据实视图的动态选择[J]. 软件学报, 2002, 13(6): 1090 - 1096.
- [7] 杨少军, 范金存, 李庆忠. 数据仓库中物化视图的选择[J]. 计算机应用, 2003, 23(9): 58 - 60.
- [8] 严勇, 袁晴晴, 周皓峰, 等. 物化视图选择方法的研究[J]. 计算机科学, 2003, 30(10): 251 - 254.
- [9] 祁文文, 徐彬, 谭红星. 数据立方体中实视图的选择[J]. 河南大学学报: 自然科学版, 2001, 31(1): 20 - 25.