数据仓库中物化视图选择算法的分析和比较①

林　巧

(浙江师范大学 数理与信息工程学院，金华 321004)

摘　要：物化视图的选择一直是数据仓库领域的研究热点。介绍了目前存在的多种典型的静态和动态选择算法，对各种算法的性能、时间复杂度等进行了分析和比较，并给出了一个优化的物化视图选择算法，最后还分析了多种混合选择方法，指出该方法是物化视图选择问题的一个新的研究方向。

关键词：数据仓库；物化视图；选择算法；静态；动态

Analysis and Comparison on Selection Algorithms of Materialized View in Data Warehouse

LIN Qiao

(College of Mathematics Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China)

Abstract: The selection of materialized view has always been a research hotspot in data warehouse domain. Some representative static and dynamic selection algorithms on the current are introduced. Performance and time complexity of these algorithms are analyzed and compared, then an optimization selection algorithm of materialized view is given. Finally some hybrid selection method that to be a new research direction of materialized view selection problem are analyzed.

Key words: data warehouse; materialized view; selection algorithms; static; dynamic

1　引言

数据仓库[1]是面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理人员的决策。经过多年发展，数据仓库已广泛应用于各行业，随着时间的推移，数据仓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，并已成为数据仓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，所以如何选择一组合适的视图进行物化就成为数据仓库查询中的一个重要问题。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[2]。该问题已被证明为 NP-完全问题，其最优解的复杂度是 $O(2^n)$，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。

十几年以来，很多研究人员对于物化视图选择与管理问题进行了大量而深入的研究，其中斯坦福、维斯康辛等大学计算机系、IBM 的 Almaden 研究中心以及微软和 AT&T 等机构都成立了专门的研究小组，专门从事这方面的研究。国内的国防科技大学、中国人民大学、华中科技大学、复旦大学以及香港科技大学等院校这方面的研究也较为活跃[3]。本文根据视图维护的频率，将物化视图选择算法分成三大类：静态、动态及混合选择算法，分别介绍了目前存在的多种典型的静态和动态选择算法，对各种算法的性能、时间复杂度等进行了分析和比较，并给出了一个优化的物化视图选择算法，最后还分析了多种混合选择方法，这种方法是物化视图选择问题的一个新的研究方向。

① 收稿时间:2011-09-07;收到修改稿时间:2011-11-10

2 物化视图的静态选择算法

在静态选择方法中，系统会在维护时间窗口内，根据查询的统计数据，把那些比较频繁发生的查询进行物化，从而提高以后到达的查询的响应速度，在下一个维护时间窗口到来之前，这些物化视图不发生变化[4]。目前已有许多关于物化视图静态选择算法的研究，下面介绍几种比较典型的静态选择算法，分析和比较这些算法的性能并归纳静态选择算法的缺点。

2.1 基于多维数据格的 BPUS 算法和 PBS 算法

1996 年，斯坦福大学的 Harinarayan 等人首先提出了多维数据格的概念，并给出了基于多维数据格的视图选择贪婪算法 BPUS(Benefit Per Unit Space)[5]，证明了该算法可以取得和其他多项式时间算法一样好的性能。虽然 BPUS 比穷尽搜索要快，但是维斯康辛大学的 Shukla 等人通过大量实验发现，BPUS 算法时间复杂度较高，算法的时间会随着维数的增加而飞速增加。为此，Shukla 等人于 1998 年提出了一种基于多维数据格的、简单快速的 PBS(Pick By Size)[6]算法来选择物化视图，其运行速度要比 BPUS 快几个数量级，它继承了 BPUS 的线性代价模型，这两种选择策略分析比较如表 1 所示。

表 1 BPUS 算法与 PBS 算法的比较

	BPUS	PBS
视图选择标准	单位空间收益的降序	聚集视图大小的升序
响应时间曲线	生成很慢	生成很快
时间复杂度	$O(kn^2)$	O(nlogn)

PBS 算法的复杂度虽然有了极大的改善，但其运行时间在多维数据集维的数量上仍是指数级增长，因而在实用性上依然没有根本的改进。

2.2 基于随机优化的遗传算法和模拟退火算法

随机优化算法以统计学理论为基础，使用评估函数来指导搜索进程不断向最优解逼近。遗传算法是随机算法的一种形式，它对于复杂的优化问题无需建模和进行复杂运算，只需要利用遗传算法的算子就能寻找到问题的最优解或次优解。采用遗传算法来解决视图选择问题基于的原理是：数据仓库中存在大量的视图和查询，并且是不断变化的。因此，视图选择问题就可以看成是一个具有大量状态空间的复杂问题。当采用遗传算法来解决视图选择问题时，需要考虑两个问题[7]：一个是表示方法问题，即如何把视图选择问题表示成遗传算法可以理解的内容；另一个问题是，遗传算法中的字符串只要发生一点小变化，就会使视图选择产生非常大的变动，有时这种变动会产生无效的结果，因此就需要采取有效的方法剔除无效的结果。

文献[8]最先提出用遗传算法解决视图选择问题，以后的许多相关研究都与该文献大同小异。在遗传算法中，不可行解的存在是一个很大的问题，文献[9]引入惩罚函数解决了这个问题。文献中指出，对于视图选择问题，遗传算法本身并不保证能够获得好的近似最优解，该方法的效果好坏取决于很多因素，比如正确的问题定义、算法的设置以及繁琐的算法参数调整。

与采用遗传算法解决视图选择问题类似，文献[10]提出采用模拟退火算法来解决视图选择问题，该算法是在 1982 年由 Kirkpatrick 等人首次提出，是一种基于 Monte Carlo 迭代求解法的启发式随机搜索算法。模拟退火算法不只是接受优化的解，也会以一定的概率接受非优化的解，这个算法可以找到全局最优解，并且只需搜索一部分状态空间。由于模拟退火算法的快速性，所以可以用于维数较大的物化视图选择问题中。但是，利用模拟算法进行视图选择的效果取决于选择正确的参数，需要通过多次实际测试得到最好的效果。

遗传算法和模拟退火算法的实现目标是在存储空间约束的条件下，取得较好的查询性能和较低的视图维护代价。

2.3 物化视图选择的优化算法设计

为了使物化视图的存储开销和查询时间开销的和最小，本文设计了一个基于遗传算法的物化视图优化算法，算法描述如下：

Step1：输入物化视图优化信息表；

Step2：根据位串长度（视图个数）确定种群规模；

Step3：i=1；

Step4：如果 i<=种群规模，反复执行 Step5~Step8，直到 i>种群规模，执行 Step9；

Step5：随机初始化视图位串，得到一个初始个体 B；

Step6：B 加入初始种群 P(0)；

Step7：用目标函数计算个体适应值；

Step8：i++；

Step9：gen=0；

Step10：如果不满足优化准则，反复执行 Step11~Step14，直到满足优化准则，执行 Step15；

Step11: 按适应值大小和选择策略从 P(gen)中选择再生个体；

Step12: 按交叉概率 P_c 和变异概率 P_m，对再生个体执行交叉、变异操作，生成新一代种群 P(gen+1)；

Step13: 用目标函数计算 P(gen+1)中个体适应值；

Step14: gen=gen+1；

Step15: 用 P(gen)中的最好个体作为系统的视图物化选择方案。

这个优化算法以查询的时间开销和物化视图的存储开销作为衡量标准建立代价估算模型，通过实验证明该算法是可行的，它可使系统在物化视图方面的存储开销和查询时间开销的和最小。

2.4 其他静态选择算法

除了以上介绍的静态选择算法，还有基于 MVPP（Multiple View Processing Plan）的方法，是 Yang 和 Karlapalem 等人于 1997 年在文献[11]中提出来的一种组织和表示查询的方法。该方法重点关注的是在最小化查询代价时的多查询优化，虽然也考虑视图维护代价，但却忽略了视图维护过程的优化问题，只是把重新计算作为视图更新策略。因此，为了弥补这个不足，文献[12]引入了视图维护策略，同时考虑了多查询优化和视图维护过程优化问题。

Gupta 于 1997 年在文献[13]中提出了 AND-OR 图的概念，并给出了构建 AND-OR 图的方法。文献中对于不考虑更新代价的 AND 图，分别给出了 Greedy 算法和 Greedy-interchange 算法；对于考虑更新代价的 AND 图，重点讨论了物化视图集合中的各个视图的更新频率小于查询频率时的情形。为解决维护代价约束下物化视图的选择问题，文献[14]在 OR 图和 AND-OR 图上提出了反向树贪婪算法和 A*启发算法。文献[15]提出了两相贪婪和集成贪婪算法。然而在实际应用中，特别是当数据仓库的数据维数超过 10 时，这些算法效率往往得不到有效保证。

2.5 静态选择算法的缺点

物化视图的静态选择方法确实改善了数据仓库的总体查询性能，但是它违背了联机分析处理 OLAP(On-Line Analysis Processing)和决策支持系统 DSS (Decision Support System)的动态本质。首先，用户查询的模式内容很难预料，对特别查询的支持程度差；其次，当数据仓库中的数据和查询的特征随着时间发生变化时，所选择的物化视图集可能很快就过时了；再次，系统没有办法改变一个错误的选择结果，更无法利用那些不能被物化视图集合回答的查询的中间结果[7]。

3 物化视图的动态选择算法

随着数据仓库系统的运行，用户查询请求的动态变化会导致物化视图集的一部分视图收益下降，部分未被物化的视图收益上升，使得物化视图集的总收益下降，所以必须通过动态选择算法加以解决。动态选择算法在查询过程中，根据查询类型的分布动态选择视图物化，克服了以上静态选择算法的缺点。目前，研究人员已经提出不少动态选择算法，下面介绍和分析几种比较典型的动态选择算法。

3.1 基于单位空间上查询频率的 FPUS 算法

为了满足动态查询，文献[16]在对静态选择算法改进的基础上，提出了基于单位空间上查询频率的即时调整的算法 FPUS（Frequency Per Unit Space），即不要求查询分布情况已知，也不需要假设查询均匀分布，而是根据收集到的查询分布情况，对物化视图进行动态调整，具有一定的创新性。

但 FPUS 算法没考虑视图间的依赖关系，忽略了物化视图的维护时间，在每次查询后都要进行全体物化效益的比较，特别是对于查询密度很高的情况不能适应。因此，仍未有效地提高系统的效率，而且该算法的即时调整策略会导致物化视图集频繁的“抖动”，使物化集缺乏稳定性，也将使很多经过优化的查询方案和优化路径不能重复利用，反而在一定程度上增加了查询开销，从而使该算法失去真正的实用价值。

3.2 基于缓存预测的 PROMISE 算法

Sapia 于 2000 年在文献[17]中提出的 PROMISE 方法是基于当前的查询来预测未来查询的结构和尺寸，该方法的原理与缓存预抓取原理相同。缓存预抓取技术已经被广泛地加以研究，但是不能把这些技术直接应用到 OLAP 环境中。要想支持缓存预抓取策略，查询负载必须具有可导航性，即连续几个查询之间具有相关性，并且前后两个查询必须大于预抓取某个对象所消耗的时间。在 OLAP 系统中，如果用户在前一个分析结果的基础上继续得到下一个分析结果，那么 OLAP 查询负载就具有可导航性。通过实验分析验证了缓存预抓取策略对 OLAP 应用的可行性以后，该文献引入了马尔可夫模型，分别设计了结构预测模型和

基于值的预测模型，以对用户的查询行为进行预测。

PROMISE 方法也存在缺点。在实际应用中，可能的查询数量通常比较大，要预测下一个阶段是哪个查询是非常耗时的，而且单个物化视图的粒度需要足够小，才能捕捉到查询之间细微的差别。然而，物化视图的粒度越细，更新代价就越高。

3.3 其他动态选择算法

2007 年冯少容等学者根据用户查询多样性的特点，提出了基于粗糙集聚类的物化视图的动态调整算法 RSCDMV[18]，该算法在对物化视图进行粗糙集聚类的基础上，既满足了用户查询多样性需求，而且兼顾了维的层次关系因素。2008 年，张东站等学者提出了一种解决物化视图动态选择新策略 NDSMMV[19]，包括候选视图生成算法、物化视图调整算法等一系列子算法。该算法的主要思想是先筛选出部分候选视图，然后通过定时地判断查询分布情况是否发生变化，来决定何时对物化视图进行动态调整，避免了物化视图集的“抖动”。2009 年，文献[20]提出了一种数据仓库中基于聚类的动态物化视图选择算法 CBD-MVS，该算法采用层次聚类技术对用户查询语句进行聚类，提出视图合并算法建立候选物化视图，利用 BPUS 算法生成最终应该被物化的视图。由于算法采用聚类技术，所以实现了完全的动态化。

4 混合选择算法

关于物化视图的静态选择算法和动态选择算法还有不少，限于篇幅，在此就不一一介绍了。通过以上的分析和比较可知，静态选择算法可以提高系统整体查询性能，但却无法适应动态变化的查询。动态方法可以提高视图选择算法适应性，但是，过多的视图调整也会加大系统的开销。因此，不少混合选择算法相继被提出。

文献[21]提出了一种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的混合算法。高层算法根据查询的局部处理方案搜索良好的全局处理方案集，采用启发式算法；低层算法根据特定的全局处理方案选择具有最小总开销的最优物化视图集，采用遗传算法。

结果表明，混合算法比采用单纯的启发式算法得到的解要好，尤其在查询数量少的时候，优势更加明显，实现了高质量的解和低计算开销的目标。

文献[22]针对传统遗传算法的缺点，提出了一种结合遗传算法和模拟退火算法思想的混合算法。该算法的实现目标是在维护代价限制条件下寻找总查询代价最小的视图集。

实验表明，混合遗传算法可得到比传统遗传算法适应度更高，即总查询代价更小的解，混合遗传算法在解决物化视图选择问题时的确优于传统遗传算法。该混合算法有效的解决了传统遗传算法“过早收敛”的通病，为物化视图的选择问题提供了一种新的有效方法。

文献[23]将蚁群算法和遗传算法相结合用于物化视图选择问题。该算法的实现目标在存储空间的限制条件下，选择一组合适的视图进行物化，使得查询集合总的查询代价最小。

利用遗传算法较强的全局搜索能力对蚂蚁每次的搜索结果进行优化改良，并在信息素更新时，同时考虑最优、最差路径上的信息素更新。实验结果表明，该算法不仅提高了解的收敛速度，也成功解决了蚁群算法易“早熟”而引起的停滞现象。

文献[24]提出了一种静态选择方法和动态选择方法有效结合的混合选择方法，既能充分利用静态选择方法改善查询响应时间的作用，又能发挥动态方法自动视图调整的功能。该方法的主要思想是，把视图集合划分为动态视图集合和静态视图集合，从动态视图集合中选出的物化视图可以即时生成或被替换，而从静态视图集合中选出的物化视图可以保留好几个视图维护窗口。

5 结语

物化视图是提高数据仓库查询响应能力的重要方法，物化视图的选择一直是数据仓库领域的研究热点。本文对物化视图的选择算法进行了分类介绍，分析和比较了每类算法的代表性研究成果，并给出了一种基于遗传算法的物化视图优化算法。通过讨论可知，单一的静态选择算法或单一的动态选择算法都有其各自的优点和局限性，而混合方法可以综合利用静态方法和动态方法的优点，并尽量避免二者的缺点，可以取得比单一的静态或动态方法更好的效果。但是，目前混合方法的研究比较少，相信今后会有更多的这方面研究成果出现，混合方法的研究已成为物化视图选择问题的一个新的研究方向。

参考文献

1 Inrnon WH. Building the data warehouse. 4th ed. New York: Wiley, 2005.

2 林小静,薛永生.数据仓库中物化视图选择策略.计算机工程与设计, 2007,28(13):3056−3059.

3 朱文,毛琴辉,薛燕.数据仓库中物化视图维护算法的分析和比较.现代计算机, 2008,281(4):58−60.

4 Choi CH, Yu JX, Lu HJ. Dynamic materialized view management based on Predicates. In: Zhou XF, Zhang YC, Orlowskal ME, eds. Proc. of the 5th Asia-Pacific Web Conf. on Web Technologies and Applications (APWeb 2003). Xi'an: Springer−Verlag, 2003:583−594.

5 Harinarayan V, Rajaraman A, Ullman JD. Implementing data cubes efficiency. Proc. of ACM SIGMOD Int'l Conf on Management of Data. New York: ACM Press, 1996:205−227.

6 Shukla A, Deshpande P, Naughton JF. Materialized view selection for multidimensional datasets. Proc. of VLDB, 1998:488−499.

7 林子雨,杨冬青,王腾蛟.实视图选择研究.软件学报, 2009, 20(2):193−213.

8 Zhang C, Yang J. Genetic algorithm for materialized viewselection in data warehouse environments. In: Mohania MK, Tjoa AM, eds. Proc. of the 8th Int'l Conf. on Data Warehousing and Knowledge Discovery (DawaK'99). Florence: Springer-Verlag, 1999:116−125.

9 Lee M, Hammer J. Speeding up materialized view select-ion in data warehouses using a randomized algorithm. Int'l Journal of Coeive Information Systems, 2001,10(3):327−353.

10 Derakhshan R, Dehne F, Korm O, Stantic B. Simulated annealing for materialized view selection in data warehouse environments. In: Hamza MH, eds. Proc. of the 24th IASTED Int'l Conf. on Database and Applications. Innsbruck: IASTED/ACTA Press, 2006:8−94.

11 Yang J, Karlapalem K, Li Q. Algorithm for materialized view design in data warehousing environment. In: Jarke M, Carey MJ, Dittrich KR, eds. Proc. of the 23rd lnt'l Conf. on Very Large Data Bases (VLDB'97). Athens: Morgan Kaufmann Publishers, 1997:136−145.

12 Yousri NA, Ahmed KM, EI-Makky NM. Algorithms for selecting materialized views in a data warehouse. Proc. of the 3rd ACS/IEEE Int'1 Conf. on Digital object Identifier (AICCSA2005). Cario: IEEE Computer Society, 2005: 27−35.

13 Gupta H. Selection of views to materialize in a data warehouse. In: Afrati FN, Kolaitis PG, eds. Proc. of the 6th Int'1 Conf. on Database Theory (ICDT'97). Delphi: Springer-Verlag, 1997:98−112.

14 Gupta H, Mumick I S. Selection of views to materialize in a data warehouse. IEEE Trans. on Knowledge and Data Engineering, 2005,17(1):24−43.

15 Gupta H, Harinarayan V, Rajaraman A, et al. Index selection for OLAP. Proc of the 13th Int'l Conf on Data Engineering. Birmingham: IEEE Computer Society Press, 1997:208−219.

16 谭红星,周龙骧,多维数据实视图的动态选择.软件学报, 2002,13(6):1090−1096.

17 Sapia C. PROMISE: Predicting query behavior to enable predictive caching strategies for OLAP systems. In: Kambayashi Y, Mohania MK, Tjoa AM, eds. Proc. of the 2nd Int'1 Conf. on Data Warehousing and Knowledge Discovery (DaWaK2000). London: Springer−Verlag, 2000:224−233.

18 冯少荣,肖文俊.基于粗糙集聚类的物化视图动态调整算法.计算机工程, 2007,33(23):185−188.

19 张东站,黄宗毅,薛永生.NDSMMV——一种多维数据集物化视图动态选择新策略.计算机研究与发展, 2008, 45(5):901−908.

20 吕晓,陈耿,朱玉全.基于聚类的动态物化视图选择研究.计算机工程与设计, 2009,30(15):3638−3640.

21 张晓辉,袁愿,虞健飞.数据仓库物化视图选择的混合算法.计算机应用, 2003,23(7):92−97.

22 徐海涛,郑宁.数据仓库中物化视图选择的一种混合算法.计算机工程与设计, 2005,26(10):2752−2755.

23 龚安,窦万蕊,王彦.基于蚁群-遗传算法的物化视图选取策略.微计算机应用, 2010,31(1):15−20.

24 Shah B, Ramachandran K, Raghavan V, Gupta H. A hybrid approach for data warehouse view selection. Int'1 Journal of Data Warehousing and Mining, 2006,2(2):l−37.