第 33 卷 第 23 期 计 算 机 工 程 2007 年 12 月
Vol.33 No.23 Computer Engineering December 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)23—0185—04 文献标识码：A 中图分类号：TP311.13

基于粗糙集聚类的物化视图动态调整算法

冯少荣[1,2]，肖文俊[1]
(1. 华南理工大学计算机科学与工程学院，广州 510640；2. 厦门大学信息科学与技术学院，厦门 361005)

摘 要：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)。该算法在对物化视图进行粗糙集聚类的基础上进行动态调整，这不仅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。实验结果证明，随着用户查询集合的增大，查询集的动态性和多样性更加明显，因此，RSCDMV 算法更具有优势。

关键词：物化视图；动态调整；粗糙集；算法

Dynamic Materialized View Algorithm Based on Rough Set Clustering

FENG Shao-rong[1,2], XIAO Wen-jun[1]
(1. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640;
2. College of Information Science and Technology, Xiamen University, Xiamen 361005)

【Abstract】 Because of user's various inquires, a new algorithm, named rough set clustering-based dynamic materialized view algorithm(RSCDMV) is presented. Based on rough set clustering on materialized view, the algorithm can execute dynamic adjustment which both satisfies the variety of the queries and take the hierarchy of dimension into consideration. Experimental results show, as the queries set increase, RSCDMV will show more advantages as inquires change.

【Key words】 materialized view; dynamic adjustment; rough set; algorithm

物化视图是指对一部分视图预先进行计算并加以物理存储，这样可以有效加快数据仓库对查询的响应速度。由于空间的限制，物化视图如何选择以最大限度地提高总体查询的效率已成为数据仓库领域的研究热点。而物化视图的选择是 NP-hard 问题[1]。

文献[1]提出了基于格模型的贪婪算法 BPUS；在此基础上，文献[2]讨论了带有 B-树索引的物化视图的选择问题；文献[3]提出了以物化视图的尺寸为选择标准的 PBS 算法；文献[4]提出了筛选候选视图的方法，以提高选择的效率。这些方案均基于查询分布已知，或查询均匀分布的前提下。而实际 OLAP 系统中的查询是随机的，均匀分布的假设常常不能成立，所以，这些静态算法具有局限性。

为了满足动态查询，文献[5]提出了基于单位空间上的查询频率的视图动态选择方法 FPUS，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。但 FPUS 没考虑视图间的依赖关系，且忽略了物化视图的维护时间，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”。

1 问题概述

现有动态调整方案的淘汰算法通常是：在物化视图中，根据一定的评价标准，将被认为不符合要求的物化视图淘汰。这种实现方法虽然简单，但是破坏了物化视图的多样性，有可能将上文提到的使用频率低，但依然重要的物化视图淘汰，从而导致某类查询的低响应度。为了解决上述问题，可以利用粗糙集理论[6~7]，即先将物化视图集合进行聚类，当执行动态调整算法时，判断该新物化视图属于哪个类别，而后只针对该类别对应的物化视图执行淘汰算法，并将新物化视图归入该类集合中。若新物化视图不属于现有任何类别，则按传统淘汰算法，将物化视图集合中最不符合要求的物化视图淘汰，而新物化视图作为新类别插入物化视图集合中。

2 基于粗糙集的物化视图聚类

文献[8]提出了一种概括的基于粗糙集的层次聚类算法，本文根据物化视图的特殊性，提出了基于粗糙集的物化视图聚类算法 (rough set-based clustering on materialized view, RSCMV)。为此，下面给出如下相关概念：

定义 1 (知识表示系统) 称 $S=(U,A,\{V_a\},f)$ 为知识表示系统，其中，U 为非空有限集，称属性集合；V_a 为属性 $a\in A$ 的值域；$f:U\to V_a$ 为一单射，使论域 U 中任一元素取属性 a 在 V_a 中的某一唯一值。

该定义可用表格表达来实现。知识的表格表达法是一种特殊的形式语言，用符号来表达等价关系，这样的数据表称作知识表达系统 KRS。在知识表达系统数据表中，列表示属性，行表示对象，并且每行表示该对象的一条信息。根据物化视图的特点，可以将物化视图集合的信息用表格表示。

基金项目：福建省自然科学基金资助项目(A0310008)；福建省高新技术研究开放计划基金资助项目(2003H043)
作者简介：冯少荣(1964－)，男，副教授、在职博士研究生，主研方向：并行分布数据库，数据仓库，数据挖掘；肖文俊，教授、博士生导师
收稿日期：2006-12-29 **E-mail：**shaorong@xmu.edu.cn

例 1 若多维数据集 Sale_Fact 包含 4 个维属性(time, product, supplier, customer)及 1 个度量属性(sales)，并且设 Time 维有 4 个层次(年、季度、月、日)；Product 维含有 2 个层次(产品类、产品名称)；Customer 维和 Supplier 维仅有 1 个层次。在 Time 维中，定义“年”为第 4 层，“日”为第 1 层，其他依次类推。现已存在一物化视图集合，如表 1 所示，行对应物化视图，列是物化视图对应的维，表格中的数字表示对应物化视图在该维上对应的层次，数字为 0 表示该物化视图没有包含该维。由表 1 可知，物化视图 1 描述每月向每个客户所销售的每种产品的销售总量。

表 1　物化视图的表格表示

物化视图编号	Time	Product	Customer	Supplier
1	3	1	1	0
2	2	2	0	1
3	4	0	1	1
4	4	2	0	0
5	2	2	1	0

定义 2 (上近似和下近似) 假设给定知识库 $K=(U,R)$, 对每个子集 $X\in U$，定义两个 U 的子集：$\overline{X}=\cup\{Y\in U|Y\subseteq X\}$ 和 $\underline{X}=\cup\{Y\in U|Y\cap X\neq\varphi\}$分别称为上近似和下近似。

$\underline{X}$是对于知识 R、U 中所有一定能归入 X 的元素的集合，而 $\overline{X}$ 是可能归入 X 的元素的集合；$bn_R(X)$是对于知识 R 既不能归入 X 也不能归入－X 的元素的集合，称为 X 的边界域。把 $pos_R(X)=\underline{X}$ 称为 X 的 R 正域，把 $neg_R(X)=U-\overline{X}$ 称为 X 的 R 负域。

定义 3 (近似精度) 在知识库 $K=(U,R)$ 中，$X\subseteq U$ 为一非空个体集，X 的近似精度定义为

$$\theta(X)=card(\underline{X})/card(\overline{X})$$

其中，card( )表示集合 X 的基数；近似精度 $\theta(X)$ 用来反映解集合 X 的知识的完全程度。对于每一个 R 且 $X\subseteq U$，有 $0\le\theta(X)\le 1$；当 $\theta(X)=1$，X 的边界域为空，集合 X 为 R 可定义的；当 $\theta(X)<1$，则集合 X 有非空边界域，该集合为 R 不可定义的。

定义 4 (不精确集合) 设 $U=\{x_1,x_2,\cdots,x_m\}$，不精确集合 I 定义为

$$I(x_i)=\{x_j\mid d(x_i,x_j)\le\beta\}$$

其中，β 为给定的阈值；$d(x_i,x_j)$ 为 x_i 和 x_j 之间的相异度，通常它是一个非负的数值。当 x_i 与 x_j 之间越相似或“接近”，其值越接近 0，反之就越大。在基于物化视图集合的粗糙集层次聚类算法中，所有属性都是数字属性，则先将各属性值进行向量归一化处理，然后再采用欧式距离法来计算个体间的距离。

定义 5 (局部不可分辨关系) 设 $x_i\in U$，x_j 对应的局部不可分辨关系 R_i 则为以下等价类集合确定：$\{I(x_i),U-I(x_i)\}$，即 R_i 将 U 分成两个互不相交的子集。对于 $\forall x_j,x_k\in U$，它们或者属于同一子集，或者属于不同的子集。论域 U 中如有 n 个个体，则有 n 个局部不可分辨关系。

定义 6 (局部不可分辨度) x_j 和 x_k 在 R_i 中的局部不可分辨度定义为

$$\gamma_i(x_j,x_k)=\begin{cases}1, [x_j]_{R_i}=[x_k]_{R_i}\\ 0, [x_j]_{R_i}\neq[x_k]_{R_i}\end{cases}$$

其中，$[x_j]_{R_i}$、$[x_k]_{R_i}$ 分别是 x_j 和 x_k 形成的 R_i 等价类。

定义 7 (全局不可分辨度) 对于 $\forall x_j,x_k\in U$，其全局不可分辨度定义为

$$\lambda(x_j,x_k)=\frac{1}{|U|}\sum_{i=1}^{|U|}\gamma(x_j,x_k)$$

定义 8 (类的不可分辨度) 设论域 U 中的 n 个个体经过聚类后得到 $C_1,C_2,\cdots,C_r$ 共 r 个互不相交的类，类 C_p 和类 C_q 之间的不可分辨度可定义为

$$\zeta_{pq}=\frac{1}{|C_p|\times|C_q|}\sum_{x_j\in C_p,x_k\in C_q}\lambda(x_j,x_k)$$

其中，$|C_p|$ 和 $|C_q|$ 分别是类 C_p 和 C_q 的个体数。根据该定义，得到递推公式

$$\zeta_{sr}=\frac{|C_p|}{|C_r|}\zeta_{sp}+\frac{|C_q|}{|C_r|}\zeta_{sq}$$

其中，C_r 为 C_p 和 C_q 的并，其样本个数 $|C_r|=|C_p|+|C_q|$。

定义 9 (综合近似精度) 设聚类结果 $C=\{C_1,C_2\cdots,C_r\}$，这 r 个类中的每个类都对应一个近似精度，则 r 个类的综合近似精度定义为

$$\alpha(C)=\sum_{i=1}^{r}\theta(C_i)\log(\theta(C_i))$$

本文采用的是凝聚的层次聚类算法，其中，在定义 4 中，很难预先确定 β 值，所以，需要自动调整 β 值以获得更优的解。在本算法中，采用 β 值为 0.005～0.2，步长为 0.001。不同的 β 值对应不同的不精确函数，从而得到不同的聚类结果。在下面描述的算法中，采用综合近似精度评价聚类结果的好坏。假定，满足 max($\alpha(C)$) 的聚类结果为较优解，对应的 β 值就是最终解。RSCMV 算法描述如下：

输入 N 个物化视图样本 X={x_1,x_2,…,x_n}

输出 被标记聚类类别的 N 个物化视图样本

(1)初始化：$\max\alpha=-1;\beta=0.005$;

(2)for i=1 to N-1

　　for j=i to N

　　根据欧氏距离公式计算 $d(x_i, x_j)$;

(3)for i=1 to N

　　计算 $I(x_i)$和 R_i;

(4)for i=1 to N

　　for j=1 to N-1

　　for k=j+1 to N

　　计算 $r_i(x_j, x_k)$;

(5)for j=1 to N-1

　　for k=j+1 to N

　　计算 $\lambda(x_j, x_k)$;

(6)将 N 个物化视图样本各归成一类，即 C_i={x_i}, i=1, 2,…,n;

(7)for i=1 to N-1

　　for j=i+1 to N

　　计算 ζ_{ij};

(8)找出不可分辨度最大的两个类 C_p,C_q，如果 ζ_{pq} 小于所设的阈值，则跳转至(11);

(9)将 C_p,C_q 合并成一个新类 C_r，并计算新类与其他类之间的不可分辨度;

(10)如果剩下不止一个类，则跳转至(8);

(11)计算此次得到聚类结果 C 的综合近似精度 $a(C)$;

如果 $a(C)$>max a 则 max a=$a(C)$;

(12) β=β+0.001;

(13)如果 β≤0.2，则跳转至(2);

(14)选取 max a 所对应的聚类结果为最终的聚类结果。

步骤(2)、步骤(5)和步骤(7)的时间复杂度为 $O(n^2)$，步骤

(3)的时间复杂度为 $O(n)$，步骤(4)的时间复杂度为 $O(n^3)$，步骤(8)～步骤(10)是一个循环，时间复杂度为 $O(n^3)$，如果引入排序机制时间复杂度可以降到 $O(n^2\log n)$。故该算法的时间复杂度为 $O(n^3)$，空间复杂度为 $O(n^2)$。

3 基于粗糙集聚类的物化视图动态调整算法

基于粗糙集聚类的物化视图的动态调整算法(rough set clustering-based dynamic materialized view algorithm, RSCDMV)具体实现如下:

输入 被标记聚类类别的物化视图集合 M_0；用户查询集 Q 所对应的视图集合 M_1；可用空间 space；

输出 调整后的物化视图集合 M_0

(1)初始化: $M_{add}=\Phi$; $M_{remove}=\Phi$; $M'=M_1-M_0$;

(2)将 M'中各个视图样本看成单独类，并根据定义 8 中的公式计算它们与其他类的不可分辨度;

(3)while $M'\neq\Phi$ {

在 M'中找到 $B(C_i, M_0\cup\{C_i\})/|C_i|$最大的 C_i ;

if $space\geq|C_i|$ then{

$M_{add}=M_{add}\cup\{C_i\}$;

$space=space-|C_i|$;

找出与 C_i 之间的不可分辨度最大的类，并将 C_i 标记成该类;}

else{

找出 C_i 与其他类的最大不可分辨度 ζ_{max},

if $\zeta_{max}<\eta$then

找当前物化视图集合中找出 $B(C_j,M_0)/|C_j|$最小的物化视图 C_j;

else

找出该类中 $B(C_j,M_0)/|C_j|$最小的物化视图 C_j;

if $B(C_i,M_0\cup\{C_i\})/|C_i|\geq B(C_j,M_0)/|C_j|$ then

if $space+|C_j|\geq|C_i|$ then{

$M_{remove}=M_{remove}\cup\{C_j\}$;

$M_0=M_0-\{C_j\}$;

$M_{add}=M_{add}\cup\{C_i\}$;

$M_0=M_0\cup\{C_i\}$;

if$\zeta_{max}<\eta$then

将 C_i 单独标记成一类;

else

将 C_i 标记成在 C_i 与其他类的最大不可分辨度 ζ_{max} 所对应的类;

$space=space+|C_j|-|C_i|$;

$M'=M'-\{C_i\}$；}

else{

$M_{remove}=M_{remove}-\{C_j\}$;

$M_0=M_0-\{C_j\}$;

$space=space+|C_j|$；} }

(4)for M_{remove} 中每个物化视图 C_k

删除物化视图 C_k

(5)for M_{add} 中每个视图 C_r

物化 C_r;

其中，η 是较小的阈值，若视图与其他类的最大不可分辨度 $\zeta_{max}<\eta$，则表示该视图不满足归入现有类的要求，即无法归入现有类中，则它可以单独成为一个新类。

若可用空间 space 大于等于新物化视图的大小，则直接进行物化操作，并将之归入到某一现有类中；若它与现有类的距离远不能将之归入现有类中，则将它算成一个新类；若可用空间 space 小于新物化视图的大小时，需要淘汰一些物化视图以腾出空间。淘汰规则可以分成下面两种情况：(1)若新视图能归入其中一个现有类中，则淘汰该类中单位空间效益最低的物化视图；(2)若新视图无法被归入一个新类中，淘汰所有物化视图中单位空间效益最低的那些视图。

物化视图之间的联系以及区别，很大程度上可以表现在维及维层次上，而在维层次结构上，现有的动态调整算法并没有太大的体现。而本文所提出的方案中，在执行基于粗糙集聚类的物化视图动态调整算法之前，会调用基于粗糙集的物化视图聚类算法，不仅满足了用户查询多样化的要求，而且在聚类操作中，物化视图的信息表格描述的正是视图对应的维层次关系，所以，聚类得出的结果也更能体现视图的维层次结构。

在系统的运行过程中，需要一个监听器不断监听系统对于用户查询的响应度，这是监听器通过计算最近发生的 K 个查询而得到的。若当前的查询响应度低于系统所设定的阈值时，监听器就会去触发系统去调用接下来所描述的物化视图的动态调整算法。这样也可以避免文献[5]中因为即时调整而导致物化视图集合频繁的“抖动”。

4 实验结果

测试环境中的硬件平台为 DELL OPTIPLEX GX270(P4 2.40GHz CPU，512MB RAM)，运行的操作系统平台是 Windows 2003 Server，数据库平台是 Orcacle 9i，算法由 Visual C++实现。实验中用到的数据包括 4 个维表(time, product, supplier, customer)和一个事实表(sales)。图 1 给出了该数据仓库中维表和事实表的结构。

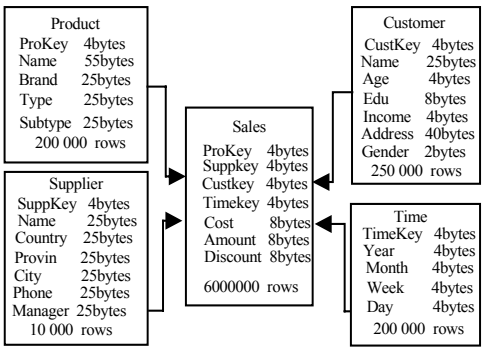

图 1 数据仓库的表结构

在文献[9]中提到，当空间限制 30%时，总代价会达到最低，所以图 2～图 4 表示的是在 30%空间的限制下，对于不同查询集大小进行物化视图选择时，查询代价、更新代价及总代价的变换情况。该实验将文献[10]中所提到的物化视图的遗传静态选择算法和本文的基于粗糙集聚类的物化视图动态调整算法相结合，并将它们所产生的结果和文献[9]所提到的 GDA 算法所产生的结果相比较。

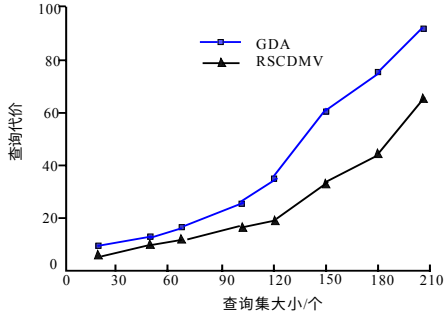

图 2 查询代价的变化情况

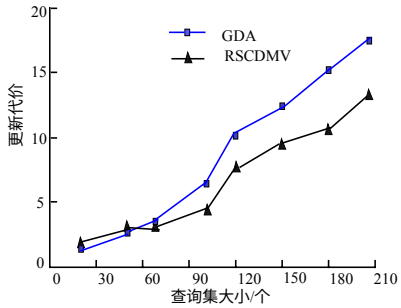

图 3　更新代价的变化情况

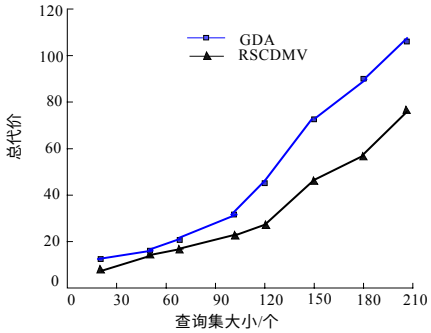

图 4　总代价的变化情况

由图 2～图 4 可知，当查询集较小时，GDA 和 RSCDMV 所耗的代价差别不大；当查询集大于 k(k=100)时，RSCDMV 算法就具有优势。因为 RSCDMV 更注重实验查询概率的分布，以及用户查询的多样性，而 GDA 只是假定用户查询集静态不变。当查询集较小时，用户查询集不会发生显著变化，当前的物化视图比较能够满足用户需求，所以，这时 GDA 和 RSCDMV 的效果大致相同；当查询集较大时，用户查询的分布情况容易发生变化，所以，如果采用基于用户查询不变上的静态选择算法，查询响应度会降低，从而导致查询代价的增加，而如果采用物化视图的动态调整算法，它会随着查询分布变化而对物化视图进行动态调整，虽然它的查询响应度也会因为查询集的增大而降低，但是它会比单纯采用静态选择算法更优化。并且，随着查询集的增大，用户查询也会趋于多样化，本文所提出的 RSCDMV 算法正是针对查询的多样化特性的，因此，随着查询集的增大，RSCDMV 算法比 GDA 更优化的效果会越来越明显。

5　结论

本文在用户查询的动态性和多样性的基础上，提出了 RSCDMV 算法，它利用粗糙集理论先对物化视图进行聚类，再在此基础上进行物化视图的动态调整。由实验结果可知，当查询集较大时可采用 RSCDMV 算法，因为，随着查询集的增大，查询分布情况发生显著变化，而且用户查询也会趋于多样化；而当查询集较小时则适合采用单纯的静态算法，因为，这时静态算法不仅简单，而且也能达到相同的效果。

参考文献

1 Harinarayan V, Rajaraman A, Ullman J D. Implementing Data Cubes Efficiently[C]//Proc. of ACM SIGMOD International Conference on Management of Data. 1996: 205-216.

2 Gupta H, Harinarayan V, Rajaraman A, et al. Index Selection for OLAP[C]//Proc. of International Conference on Data Engineering. 1997: 208-219.

3 Shukla A, Deshpande P M, Naughton J F. Materialized View Selection for Multidimensional Datasets[C]//Proc. of the 24th VLDB Conference. 1998: 488-499.

4 Baralis E, Paraboschi S, Teniente E. Materialized View Selection in a Multidimensional Dtabase[C]//Proc. of the 23rd VLDB Conference. 1999: 156-165.

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

6 Pawlak Z. Rough Sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341-356.

7 Pawlak Z, Grzymla-Busse Jetal. Rough Sets[J]. Communications of the ACM, 1995, 38(11): 88-95.

8 刘少辉, 胡　斐, 贾自艳, 等. 一种基于 Rough 集的层次聚类算法[J]. 计算机研究与发展, 2004, 41(4): 552-557.

9 Lin Wenyang, Kuo I C. OLAP Data Cubes Configuration with Genetic Algorithms[C]//Proc. of IEEE International Conference on Systems, Man, and Cybernetics. 2000: 1984-1989.

10 钟静华, 冯少荣, 段江娇, 等. 数据立方体的带有动态调整的遗传选择算法[J]. 计算机科学, 2005, 32(增刊): 445-448.

（上接第 184 页）

其中，图 5 中被圈中的是不满足集群条件的人工鱼。图 6 中被圈中的是捕食者——鲨鱼。

6　结束语

本文从复杂系统的角度研究了人工鱼群的自组织行为，建立了鱼群的多 Agent 系统模型，并在此基础上建立了鱼群的觅食、逃逸模型。但是只考虑到影响个体鱼集群的几个因素，事实上还有很多生理、心理、遗传等因素需要考虑，同时人工鱼群作为整体，还存在着个体间通信、合作捕食关系等。这些都将是笔者后续研究的内容。随着复杂系统和基于认知的人工鱼高级行为规划研究的深入，鱼群自组织行为研究的理论基础也将不断丰富，在这些理论的指导下将完善鱼群系统，生动地展现海洋鱼群的运动和生活。

参考文献

1 Reynolds C W. Flocks, Herds, and Schools: A Distributed Behavioral Model[C]//Proceedings of SIGGRAPH'87 Conference on Computer Graphics. 1987: 25-34.

2 Tu X. Artificial Animals for Computer Animation: Biomechanics, Locomotion, Perception, and Behavior[M]. [S. l.]: Springer-Verlag, 1999.

3 Shaw E. Fish in Schools[J]. Natural History, 1975, 84(8): 40-46.

4 班晓娟, 曾广平, 涂序彦. 基于自学习的人工鱼感知系统设计与实现[J]. 电子学报, 2004, 32(12).