在 VPRS 模型中保持边界的属性约简方法研究*)

孙士保[1,2]　秦克云[2]

(西南交通大学智能控制开发中心　成都 610031)[1]　(河南科技大学电子信息工程学院　洛阳 471003)[2]

摘　要　讨论了变精度粗糙集模型中现有的属性约简方法，找出了 β 约简的不足；介绍了 Inuiguchi 提出的保持决策类下近似，上近似，边界和无法预言区的属性约简定义；说明了保持下近似的属性约简就是 β 下分布约简，保持上近似属性约简就是 β 上分布约简；提出了变粗度粗糙集模型中基于边界的属性约简方法，并从理论上证明了它的正确性；最后，给出了该种方法的实现算法。经实例证明，该方法操作简单，具有很高的应用价值。

关键词　变精度粗糙集模型，属性约简，边界，算法

Research on Approximation Operator in Generalized Variable Precision Rough Set Model

SUN Shi-Bao[1,2]　QIN Ke-Yun[2]

(Intelligent Control Development Center, Southwest Jiaotong University, Chengdu 610031)[1]

(Electronic Information Engineering College, Henan University of Science and Technology, Luoyang 471003)[2]

Abstract　Attribute reduction approaches at hand in variable precision rough set model are discussed, the shortages of β-reduct are found. Definitions of attribute reduction for preserving lower approximations, upper approximations, boundary regions and unpredictable region proposed by Inuiguchi are introduced. That attribute reduction approach of preserving lower approximations in β lower distribution reduction and approach of preserving upper approximations is β upper distribution reduction are demonstrated. Attribute reduction approach boundary regions-preserved in variable precision rough set model is proposed and its correctness is proved from theoretical. Finally, the algorthm of this approach is given. The experiments show that this approach operates simply and it has better applied value.

Keywords　Variable precision rough set model, Attribute reduction, Boundary regions, Algorithm

1　引言

粗糙集(Rough Sets)理论是 Z. Pawlak 教授于 1982 年提出的一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法[1]。该理论已经成为智能计算领域的研究热点，并在信息处理、数据挖掘(DM)和数据库知识发现(KDD)等认知领域有成功的应用[2]。

粗糙集理论将分类与知识联系在一起，根据已知数据自身的不可分辨关系，通过一对近似算子，对某一给定概念进行近似表示，它是一种数据驱动的方法，本质上不需要任何关于数据和相应问题的先验知识和附加信息，因此特别适合应用于知识发现与数据挖掘领域。Pawlak 粗糙集模型的一个局限性是它所处理的分类必须是完全正确的或肯定的，因而它的分类是精确的，亦即只考虑完全"包含"和"不包含"，而没有某种程度上的"包含"和"属于"。Pawlak 粗糙集模型的另一个局限性是它所处理的对象是已知的，且从模型中得到的结论仅适用于这些对象。但在实际应用中，往往需要把从小规模对象集中得到的结论应用于大规模对象集上去。Pawlak 粗糙集模型的局限性限制了它的应用。为了克服这些局限性，Ziarko 提出了变精度粗糙集模型[3](VPRS 模型，即 Variable Precision Rough Set Model)，它是 Pawlak 粗糙集模型的扩展，它的基本思想是在 Pawlak 粗糙集模型中引入参数，即允许一定程度的错误分类率存在，它可以解决属性间无函数关系的数据分类问题，当时，变精度粗糙集就退化为 Pawlak 粗糙集。这种推广的模型有利于从看似不相关的数据中发现潜在的相关数据。目前变精度粗糙集模型已经在很多行业得到了广泛的应用[4]。

然而，基于变精度粗糙集模型的推广工作没有得到进一步的研究。到目前为止，只有 β 约简[5]、β 上(下)近似(分布)约简[6]是比较常用的。文[7]中 Inuiguchi 提出了在变精度模型中当对象的论域被分成 3 个或 3 个以上的等价类时，存在一个无法预言的区域是所有的上近似都无法覆盖到的，所以可以得到一种由决策类的下近似、决策类的上近似、决策类的边界和无法预言的区域组成的结构。在变精度粗糙集模型中，当条件属性集 $A\subseteq C$ 时不能保证 $\underline{A}(D_j)\subseteq\underline{C}(D_j)$，$\overline{A}(D_j)\supseteq\overline{C}(D_j)$，$j=1,\cdots,p$，所以 β 约简不总是保持 D_j 的下近似。例如对于 β 约简 A 来说可能一个对象属于 $\underline{C}_\beta(D_j)$ 也可能属于 $\underline{A}_\beta(D_i)(i\neq j)$，因此 Inuiguchi 提出了一组保持这种结构的约简方法，它们是：L^β-约简、U^β-约简 B^β-约简和 UN^β-约简以及由它们组合的各种约简。但遗憾的是 Inuiguchi 只是给出了保持结构的约简定义，并没有给出这些约简的方法。作者在研究中发现 L^β-约简就是文[6]中的 β 下分布约简，U^β-约简就是文[6]中的 β 上分布约简。在文[8]中证明了 β 上(下)分布约简可以保持信息系统中直辖市和不直辖市规则的一致性，所以是完全正确的约简方法。文[6]中有详细的 β 上(下)分布约简方法，所以本文不再详述。本文只讨论 Inuiguchi 提出的保持边界的 B^β-约简，并给出了一种简单的 B^β-约简方法。

2　相关概念

定义 1[7]　设 (U,A,F,D,G) 为目标信息系统，其中 (U,A,F) 为信息系统，U 为论域，即对象集，A 为条件属性集，D 为目标属性集。F 为 U 与 A 的关系集，G 为 U 与 D 的关系集。对于任意 $B\subseteq A$，记 $R_B=\{(x_i,x_j)\in U\times U: f_l(x_i)=f_l$

*)基金项目：国家自然科学基金资助项目(批准号：60474022)。孙士保　讲师，博士生，主要研究方向为智能信息处理。秦克云　教授，博士生导师，主要研究方向为逻辑与不确定性推理。

$(x_j)(\forall a_l\in B)\}$，$A(B)=U/R_B=\{[x]_B:x\in U\}$，其中$[x]_B=\{y\in U:(x,y)\in R_B\}$。同样 $R_D=\{(x_i,x_j)\in U\times U:g_l(x_i)=g_l(x_j)(\forall dl\in D)\}$，$A(D)=U/R_D=\{[x]_D:x\in U\}=\{D_1,D_2,\cdots,D_r\}$，其中$[x]_D=\{y\in U:(x,y)\in R_D\}$。$R_A=\{(x_i,x_j)\in U\times U:f_l(x_i)=f_l(x_j)(\forall a_l\in A)\}$，$U/R_A=\{[x]_A:x\in U\}=\{C_i:i\leqslant t\}$，$[x]_A=\{y\in U:(x,y)\in R_A\}$。$\beta\in[0,0.5)$。令 $\mu_{D_j}(x|B)=\frac{|[x]_B\cap D_j|}{|[x]_B|}(x\in U)$。$D_j$ 的β下近似$\underline{A}_\beta(D_j)=\{x\in U|\mu_{D_j}(x|B)\geqslant 1-\beta\}$，$D_j$ 的β上近似$\overline{A}_\beta(D_j)=\{x\in U|\mu_{D_j}(x|B)>\beta\}$，$D_j$ 的β边界域定义为$B^\beta_B(D_j)=\overline{A}_\beta(D_j)-\underline{A}_\beta(D_j)$。

定义 2 设(U,A,F,D,G)为目标信息系统，对于任意 $B\subseteq A$，B 是 B^β 边界协调集当且仅当 $BN^\beta_B(D_j)=BN^\beta_A(D_j)$，$j=1,\cdots,r$。

定义 3 B^β 约简：设(U,A,F,D,G)为目标信息系统，条件属性集 $C\subseteq A$ 叫做 B^β 约简当且仅当满足：①$BN^\beta_C(D_j)=BN^\beta_A(D_j)$，$j=1,\cdots,r$ 和②不存在 $B\subseteq C$，使 $BN^\beta_B(D_j)=BN^\beta_A(D_j)$，$j=1,\cdots,r$。

定理 1 设(U,A,F,D,G)为目标信息系统，其中(U,A,F)为信息系统，U 为论域，即对象集，A 为条件属性，D 为目标属性集。F 为 U 与 A 的关系集，G 为 U 与 D 的关系集。对于任意 $B\subseteq A$，记

$$BN^\beta_B(x)=\{D_j:x\in BN^\beta_B(D_j)\}(x\in U)$$

则 B 是 B^β 边界协调集当仅当 $BN^\beta_B(x)=BN^\beta_A(x)(x\in U)$。

证明：利用任意 $x\in BN^\beta_B(D_j)$ iff $D_j\in BN^\beta_B(x)$，并且 $x\in BN^\beta_B(D_j)$ iff $D_j\in BN^\beta_A(x)$。证毕。

定理 2(知识约简的判定定理) B 是 B^β 边界协调集当仅当 $\forall x,y\in U$，当 $BN^\beta_B(x)\neq BN^\beta_A(y)$ 时，有$[x]_B\cap[y]_B=\varnothing$。

证明：记 $J([x]_B)=\{[y]_A:[y]_A\subseteq[x]_B\}$。由 $B\subseteq A$ 可知 $J([x]_B)$构成了$[x]_B$ 的分划。

设 B 是 B^β 边界协调集。对于 $x,y\in U$，若$[x]_B\cap[y]_B\neq\varnothing$，则$[x]_B=[y]_B$，于是 $BN^\beta_B(x)=BN^\beta_B(y)$。由于 B 是 B^β 边界协调集，由定理 1 知 $BN^\beta_B(x)=BN^\beta_A(x)$ 且 $BN^\beta_B(y)=BN^\beta_A(y)$，从而 $BN^\beta_A(x)=BN^\beta_A(y)$。因而当 $BN^\beta_A(x)\neq BN^\beta_A(y)$ 时，就有$[x]_B\cap[y]_B=\varnothing$。反之，对于 $\forall x\in U$，当$[y]_A\subseteq[x]_B$ 时，显然$[x]_B\cap[y]_B\neq\varnothing$，从而由假设得 $BN^\beta_A(x)=BN^\beta_A(y)$。

对于 $\forall j\leqslant r$，若 $x\in BN^\beta_B(D_j)=\overline{R}^\beta_B(D_j)-\underline{R}^\beta_B(D_j)$，则 $x\in\overline{R}^\beta_B(D_j)$且 $x\notin\underline{R}^\beta_B(D_j)$，所以$[x]_B\subseteq\overline{R}^\beta_B(D_j)$且$[x]_B\not\subset\underline{R}^\beta_B(D_j)$，则$[x]_B\subseteq BN^\beta_B(D_j)$。由于$[x]_B=\cup\{[y]_A:[y]_A\in J([x]_B)\}$，故对任意的$[y_0]_A\in J([x]_B)$，有$[y_0]\subseteq BN^\beta_B(D_j)$。由$[x]_B\subseteq BN^\beta_B(D_j)$得 $1-\beta<D(D_j/[x]_B)<\beta$，同样由$[y_0]\subseteq BN^\beta_B(D_j)$有 $1-\beta<D(D_j/[y_0]_A)<\beta$，因此$[y_0]\subseteq BN^\beta_A(D_j)$，所以 $y_0\in BN^\beta_A(D_j)$则 $D_j\in BN^\beta_A(y_0)$，于是 $D_j\in BN^\beta_A(x)$，因此 $x\in BN^\beta_A(D_j)$，即 $BN^\beta_B(D_j)\subseteq BN^\beta_A(D_j)$。

另一方面，若 $x\in BN^\beta_B(D_j)$，则 $D_j\in BN^\beta_A(x)$，而当$[y]_A\in J([x]_B)$时，$[x]_B\cap[y]_B\neq\varnothing$，故 $BN^\beta_A(x)=BN^\beta_A(y)$，从而 $D_j\in BN^\beta_A(y)$即 $1-\beta<D(D_j/[y]_A)<\beta$，这样就有 $D(D_j/[x]_B)=(\sum\{|[y]_A\cap D_j|:[y]_A\in J([x]_B)\})/|[x]_B|=\sum\{D(D_j/[y]_A)\cdot\frac{|[y]_A|}{|[x]_B|}:[y]_A\in J([x]_B)\}$，所以 $1-\beta<D(D_j/[x]_B)<\beta$，因此 $x\in BN^\beta_B(D_j)$，即 $BN^\beta_B(D_j)\supseteq BN^\beta_B(D_j)$。

这样便证明了 $BN^\beta_B(D_j)=BN^\beta_B(D_j)(\forall j\leqslant r)$，即 B 是 B^β 边界协调集。

定义 4 设(U,A,F,D,G)是目标信息系统，$U/R_A=\{C_i:i\leqslant t\}$记 $D^{*\beta}=\{([x]_A,[y]_A):BN^\beta_A(x)\neq BN^\beta_A(y)\}$。

用 $f_k(C_i)$表示属性 a_k 关于 C_i 中对象的取值。定义

$$D^\beta(C_i,C_j)=\begin{cases}\{a_k\in A:f_k(C_i)\neq f_k(C_j)\}, & (C_i,C_j)\in D^{*\beta},\\ A, & (C_i,C_j)\notin D^{*\beta}.\end{cases}$$

则称 $D^\beta(C_i,C_j)$为 C_i 与 C_j 的β边界可辨识属性集。称 $D^\beta=(D^\beta(C_i,C_j),i,j\leqslant t)$为目标信息系统的$\beta$边界可辨识属性矩。

定理 3 设(U,A,F,D,G)是目标信息系统，$B\subseteq A$，则 B 是 B^β 边界协调集当且仅当对于任意$(C_i,C_j)\in D^{*\beta}$，有 $B\cap D^\beta(C_i,C_j)\neq\varnothing$。

证明：设 B 是 B^β 边界协调集，对于$\forall(C_i,C_j)\in D^{*\beta}$，取 $x,y\in U$ 使 $C_i=[x]_A$，$C_j=[y]_A$，则 $BN^\beta_A(x)\neq BN^\beta_A(y)$。于是由定理 2 得$[x]_B\cap[y]_B=\varnothing$。从而存在 $a_k\in B$ 使 $f_k(x)\neq f_k(y)$，即 $f_k(C_i)\neq f_k(C_j)$，故 $a_k\in D^\beta(C_i,C_j)$，因此，$B\cap D^\beta(C_i,C_j)\neq\varnothing$。

反之，若存在$(C_i,C_j)\in D^{*\beta}$使 $B\cap D^\beta(C_i,C_j)=\varnothing$，则可取 $x,y\in U$ 使 $C_i=[x]_A$，$C_j=[y]_A$。于是一方面由$([x]_A,[y]_A)\in D^{*\beta}$知 $BN^\beta_A(x)\neq BN^\beta_A(y)$。另一方面，对于任意 $a_k\in B$，又有 $a_k\notin D^\beta(C_i,C_j)$，于是 $f_k(C_i)=f_k(C_j)$，从而 $f_k(x)=f_k(y)$，这说明$[x]_B=[y]_B$。再由定理 2 知 B 不是 B^β 边界协调集。证毕。

定义 5 设(U,A,F,D,G)为目标信息系统，$D^\beta=(D^\beta(C_i,C_j),i,j\leqslant t)$为目标信息系统的$\beta$边界可辨识属性矩阵。记

$$M^\beta=\wedge\{\vee\{a_k:a_k\in D^\beta(C_i,C_j)\}:i,j\leqslant t\}$$
$$=\wedge\{\vee\{a_k:a_k\in D^\beta(C_i,C_j)\}:(C_i,C_j)\in D^{*\beta}\}$$

则 M^β 为β边界辨识公式。

定理 4 设(U,A,F,D,G)为目标信息系统，边界辨识公式 M^β 的极小析取范式为 $M^\beta=\bigvee_{k=1}^{p}(\bigwedge_{s=1}^{q_k}a_s)$。记 $B_k=\{a_s:s=1,2,\cdots,q_k\}$，则$\{B_k:k=1,2,\cdots,r\}$是$\beta$边界约简形式的集合。

证明：对于任意 $k\leqslant p$ 和$(C_i,C_j)\in D^{*\beta}$，由极小析取范式的定义知 $B_k\cap D^\beta(C_i,C_j)\neq\varnothing$，再由定理 3 知 B_k 是 B^β 边界协调集。同时，由 M^β 知在 B_k 中去掉一个元素形成的 B'_k，则必存在$(C_i,C_j)\in D^{*\beta}$，使得 $B'_k\cap D^\beta(C_i,C_j)=\varnothing$，故 B'_k 不是 B^β 边界协调集，从而 B'_k 不是 B^β-约简。

由于β边界辨识公式中包含了所有的 $D^\beta(C_i,C_j)$，因此不存在其他的 B^β- 约简。证毕。

3 B^β-约简算法

设(U,A,F,D,G)为目标信息系统，其中(U,A,F)为信息系统，U 为论域，即对象集，A 为条件属性集，D 为目标属集。F 为 U 与 A 的关系集，G 为 U 与 D 的关系集。记 $R_A=\{(x_i,x_j)\in U\times U:f_l(x_i)=f_l(x_j)(\forall a_1)\in A)\}$，$U/R_A=\{[x]_A:x\in U\}=\{C_i:i\leqslant t\}$，$[x]_A=\{y\in U:(x,y)\in R_A\}$。同样 $R_D=\{(x_i,x_j)\in U\times U:g_l(x_i)=g_l(x_j)(\forall d_l\in D)\}$，$U/R_D=\{[x]_D:x\in U\}=\{D_1,D_2,\cdots,D_r\}$，其中$[x]_D=\{y\in U:(x,y)\in R_D\}$。条件属性集 $C\subseteq A$，$\beta\in[0,0.5]$。$\mu_{D_j}(x|C)=\frac{|[x]_C\cap D_j|}{|[x]_C|}(x\in U,i=1,\cdots,r)$。从以上分析可以得出 B^β-约简算法：

输入：设(U,A,F,D,G)为目标信息系统。

输出：目标信息系统的 B^β-约简集 B。

步骤 1. 初始化。即求 R_A，$U/R_A=\{C_1,C_2,\cdots,C_t\}$；$R_D$，$U/R_D=\{D_1,D_2,\cdots,D_r\}$。

步骤 2. 求 $\mu_{D_j}(x|A)(x\in U,j=1,\cdots,r)$，$BN^\beta_A(D_j)=\overline{A}_\beta$

$(D_j)-\underline{A}_\beta(D_j)$。

步骤 3. 求 $BN_A^\beta(x)=\{D_j: x\in BN_A^\beta(D_j)(x\in U,j=1,\cdots,r)$。

步骤 4. 求 $D^\beta=(D^\beta(C_i,C_j),i,j\leqslant t)$。其中，

$$D^\beta(C_i,C_j)=\begin{cases}\{a_k\in A: f_k(C_i)\neq f_k(C_j)\}, & (C_i,C_j)\in D^{*\beta},\\ A, & (C_i,C_j)\notin D^{*\beta}.\end{cases}$$

和 $D^{*\beta}=\{([x]_A,[y]_A): BN_A^\beta(x)\neq BN_A^\beta(y)\}$。

步骤 5. 求 β 边界辨识公式 M^β，$M^\beta=\wedge\{\vee\{a_k: a_k\in D^\beta(C_i,C_j)\}: i,j\leqslant t\}=\wedge\{\vee\{a_k: a_k\in D^\beta(C_i,C_j)\}: (C_i,C_j)\in D^{*\beta}\}$。

步骤 6. 化简 M^β，得到 M^β 的极小析取范式形式 $M^\beta=\bigvee_{k=1}^{p}(\bigwedge_{s=1}^{q_k}a_s)$，记 $B_k=\{a_s: s=1,2,\cdots,q_k\}$，则 $B=\{B_k: k=1,2,\cdots,r\}$ 是 β 边界约简形成的集合。

4 B^β-约简应用实例

文[6]中给出了一个目标信息系统，其中 $U=\{x_1,\cdots,x_6\}$ 为论域，$A=\{a_1,\cdots,a_4\}$ 为条件属性集，$D=\{d\}$ 为目标属性集。用它来求 B^β-约简，其中 $\beta=0.3$。

表 1　目标信息系统

U	a_1	a_2	a_3	a_4	d
x_1	1	0	0	0	1
x_2	0	1	1	1	2
x_3	0	1	0	0	2
x_4	0	1	1	0	2
x_5	0	1	0	0	1
x_1	0	1	0	0	1

按照 3 中的 B^β-约简算法得：

步骤 1　$U/R_A=\{C_1,C_2,\cdots,C_4\}$，其中 $C_1=\{x_1\}$，$C_2\{x_2\}$，$C_3=\{x_3,x_5,x_6\}$，$C_4=\{x_4\}$；$U/R_D=\{D_1,D_2\}$，$D_1=\{x_1,x_5,x_6\}$，$D_2=\{x_2,x_3,x_4\}$。

步骤 2　$\mu_{D_1}(x_1|A)=1$，$\mu_{D_1}(x_2|A)=0$，$\mu_{D_1}(x_3|A)=2/3$，$\mu_{D_1}(x_4|A)=0$，$\mu_{D_1}(x_5|A)=2/3$，$\mu_{D_1}(x_6|A)=2/3$，$\mu_{D_2}(x_1|A)=0$，$\mu_{D_2}(x_2|A)=1$，$\mu_{D_2}(x_3|A)=1/3$，$\mu_{D_2}(x_4|A)=1$，$\mu_{D_2}(x_5|A)=1/3$，$\mu_{D_2}(x_6|A)=1/3$。

当 $\beta=0.3$ 时，$\overline{A}_{0.3}(D_1)=\{x_1,x_3,x_5,x_6\}$，$\underline{A}_{0.3}(D_1)=\{x_1\}$，$BN_A^{0.3}(D_1)=\{x_3,x_5,x_6\}$；$\overline{A}_{0.3}(D_2)=\{x_2,x_3,x_4,x_5,x_6\}$，$\underline{A}_{0.3}(D_2)=\{x_2,x_4\}$，$BN_A^{0.3}(D_2)=\{x_3,x_5,x_6\}$。

步骤 3　$BN_A^{0.3}(x_1)=\varnothing$，$BN_A^{0.3}(x_2)=\varnothing$，$BN_A^{0.3}(x_3)=\{D_1,D_2\}$，$BN_A^{0.3}(x_4)=\varnothing$，$BN_A^{0.3}(x_5)=\{D_1,D_2\}$，$BN_A^{0.3}(x_6)=\{D_1,D_2\}$。

步骤 4　$D^{*0.3}=\{(C_1,C_2),(C_2,C_3),(C_3,C_4)\}$，$D^{0.3}(C_1,C_3)=\{a_1,a_2\}$，$D^{0.3}(C_2,C_3)=\{a_3,a_4\}$，$D^{0.3}(C_3,C_4)=\{a_3\}$。目标信息系统的 $\beta=0.3$ 边界可辨识属性矩阵的其它元素全为 A。

步骤 5　$\beta=0.3$ 边界辨识公式 $M^{0.3}=(a_1\vee a_2)\wedge(a_3\vee a_4)\wedge a_3$。

步骤 6　化简得 $M^{0.3}=(a_1\vee a_2)\wedge a_3=(a_1\wedge a_3)\vee(a_2\wedge a_3)$，从而得到 $B=\{\{a_1,a_3\},\{a_2,a_3\}\}$ 是 $\beta=0.3$ 边界约简形成的集合。

结论　在变精度粗糙集模型中，目前已有的属性约简方法并不多，该种保持边界的属性约简方法可以作为已有方法的补充。该方法条理清晰，从理论上证明了它的正确性，从实例上说明了它的可操作性，很适合用计算机编程实现。但是，能否实现复杂的大型数据库的约简还有待进一步的研究。

参 考 文 献

1 Pawlak Z. Rough sets[J]. International journal of Information & Computer Science, 1982, 11(5): 341～356
2 刘清. Rough集及Rough推理[M]. 北京：科学出版社，2003. 40～80
3 Ziarko W. Variable precision rough set model. Journal of computer system science, 1993, 46(1): 39～59
4 陶志，许宝栋，汪定伟，李冉. 基于变精度粗糙集理论的粗糙规则挖掘算法. 信息与控制，2004，33(1)：18～22
5 Beynon M. Reducts within the variable precision rough sets model: A further investigation[J]. European journal of operational research, 2001, 134: 592～605
6 Zhang Wenxiu, Liang Yi, Wu Weizhi. Information System and Knowledge Discoery[M]. Beijing: Science Press, 2003. 56～67
张文修，梁怡，吴伟志，等. 信息系统与知识发现[M]. 北京：科学出版社，2003. 56～67
7 Inuiguchi M. Seeral approaches to attribute reduction in ariable precision rough set mode[A]. In: Proceeding of Modeling Decisions for Artificial Intelligence [C]. MDAI2005, Springer-Verlag, 2005. 523～528
8 Qin Keyun, Pei Zheng, Du Weifeng. The relationship among several knowledge reduction approaches[A]. In: Proceedings of the 2nd international conference on fuzzy systems and knowledge discovery[C]. FSKD2005, ChangSha, 2005.8, Berlin: Spinger, 2005, 1: 1232～1241

(上接第 105 页)

结束语　Dwarf 是一种类树结构，通过共享前缀和收拢后缀，有效地降低了数据立方的存储开销。我们对 Dwarf 结构进行了进一步的研究，指出 Dwarf 结构中仍然存在冗余，为此提出了浓缩 Dwarf，进一步减小了 Dwarf 的存储尺寸。另外，通过将冰山立方思想融合到 Dwarf 技术之中，我们提出了冰山 Dwarf。冰山 Dwarf 适合于用户不太关心细节的应用场合，在这种情况下，冰山 Dwarf 能够极大地减小 Dwarf 的存储开销。

参 考 文 献

1 Gray J, Bosworth A, Layman A, et al. Datacube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, Sub-Totals. In: Proceedings of 12th International Conference on Data Engineering (ICDE'96). Louisiana, USA, 1996. 152～159
2 Hahn C, Warren S, London J. Edited synoptic cloud reports from ships and land stations over the globe, 1982-1991. http://cidiac. est. ornl. gov/ftp/ndp026b/SEP85L. z, 1994
3 Wang W, Feng J, Lu H, et al. Condensed Cube: An Effective Approach to Reducing Data Cube Size. In: Proceedings of 18th International Conference on Data Engineering (ICDE'02). CA, USA, 2002. 155～165
4 Vitter J S, Wang M, Iyer B. Data Cube Approximation and Histograms via Wavelets. In: Proceedings of 7th International Conference on Information and Knowledge Management (CIKM'98). Bethesda, Maryland, USA, 1998. 96～104
5 Sismanis Y, Deligiannakis A, Roussopoulos N, et al. Dwarf: Shrinking the PetaCube. In: Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'02). Madison, Wisconsin, USA, 2002. 464～475
6 Lakshmanan L V S, Pei J, Han J. Quotient Cube: How to Summarize the Semantics of a Data Cube. In: Proceedings of 24th International Conference on Very Large Data Bases (VLDB'02). Hongkong, China, 2002. 766～777
7 Lakshmanan L V S, Pei J, Zhao Y. QC-Trees: An effective Summary structure for Semantic OLAP. In :SIGMOD'03, 2003
8 Fang M, Shivkumar N, Garcia-Molina H, et al. Computing iceberg queries efficiently. In: A. Guppta, O. Shmueli, J. Widom, eds. Proceedings of 24th International Conference on Very Large Data Bases (VLDB'98). New York, USA. Morgan Kaufmann, 1998. 299～310
9 Gray J, Sundaresan P, Englert S. Quickly Generating Billion-Record Synthetic Databases. In: R. T. Snodgrass, M. Winslet, eds. Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data (SIGMOD'94). Minnesota, USA, 1994. 243～252