

P-集合与数据内搜索-应用

谢维奇^{1,2} 于秀清^{2,3}

(驻马店职业技术学院信息工程系 驻马店 463000)¹ (山东大学数学与系统科学学院 济南 250100)²
(德州学院数学系 德州 253000)³

摘 要 P-集合是把动态特性引入到有限普通集合中,改进普通集合得到的。P-集合是由内 P-集合 X^F (internal packet sets X^F) 与外 P-集合 X^F (outer packet sets X^F) 构成的集合对,或者 (X^F, X^F) 是 P-集合。利用内 P-集合,给出数据内搜索的概念,给出 $\bar{F}$ -数据的度量和依赖关系,给出 $\bar{F}$ -数据内搜索迭代算法和准则,给出数据内搜索的应用。P-集合是研究动态信息系统的一个新理论与新方法。

关键词 P-集合, $\bar{F}$ -数据, 数据内搜索, 依赖-内搜索准则, 应用

中图法分类号 O144 文献标识码 A

P-sets and Data Internal Search-application

XIE Wei-qi^{1,2} YU Xiu-qing^{2,3}

(Department of Information Engineering, Zhumadian Vocational and Technical College, Zhumadian 463000, China)¹

(School of Mathematics and System Sciences, Shandong University, Jinan 250100, China)²

(Department of Mathematics, Dezhou University, Dezhou 253000, China)³

Abstract Introduced by dynamic characteristics, P-sets can be obtained from the improved finite general sets. P-sets are a set pair which are composed of internal P-sets X^F and outer P-sets X^F , or (X^F, X^F) is P-sets. By using P-sets, the concept of data internal search, the measure and dependence relation about $\bar{F}$ -data were given, search iterative algorithm and criterion of $\bar{F}$ -data were proposed, finally the applications of data internal search were given. P-sets is a new theory and method to study dynamic information system.

Keywords P-sets, $\bar{F}$ -data, Data internal search, Dependence-internal search criterion, Applications

1 引言

文献[1,2]把动态特性引入到有限普通集合 X 中,改进有限普通集合 X ,提出 P-集合(packet sets)。P-集合是由内 P-集合 X^F (internal packet sets X^F) 与外 P-集合 X^F (outer packet sets X^F) 构成的集合对,或者 (X^F, X^F) 是 P-集合。P-集合具有动态特性。

文献[1-10]给出如下事实:有限普通集合 X , X 具有属性集合 α ,若对 α 内的属性补充, α 变成 α^F ,则集合 X 中的元素被删除, X 生成内 P-集合 X^F 。不断重复这样的过程,可以得到一个集合序列: $X_n^F \subseteq X_{n-1}^F \subseteq \dots \subseteq X_2^F \subseteq X_1^F \subseteq X$ (这里 $X_n^F \neq \emptyset$)。它们对应的属性集合满足 $\alpha \subseteq \alpha_1^F \subseteq \alpha_2^F \subseteq \dots \subseteq \alpha_{n-1}^F \subseteq \alpha_n^F$ 。如果 X_n^F 是要得到的数据,则 X_n^F 可以等价地看作在 X 内进行数据内搜索得到的。搜索过程与我们使用搜索引擎情况完全相同:在查询条件中输入关键词 α_1 ,得到相应的数据。如果增加查询关键词 α_2 ,或者查询关键词变为 $\{\alpha_1, \alpha_2\}$,就可以得到范围更小的数据集 X' , $X' \subset X$,而且 X' 是依赖 X 而存在的。不断地增加关键词,根据关键词搜索得到的数据越来越少、越来越精确,即越来越接近搜索的要求。这是一个计算机

在工程应用中司空见惯的现象。有趣的是这个现象和 P-集合(特别是内 P-集合)的动态特性十分相似。把 P-集合应用于数据搜索及所被搜索数据之间的关系和特征的讨论鲜有人给出。在能见到的 P-集合研究论文中,找不到与本文相似的研究。

本文利用 P-集合与它的动态特性,研究数据内搜索的过程和特征,给出 $\bar{F}$ -数据的概念、 $\bar{F}$ -数据的度量、 $\bar{F}$ -数据的依赖关系、 $\bar{F}$ -数据内搜索迭代算法和准则、数据内搜索的应用。

为了便于讨论,容易接受本文的结果,把 P-集合与它的结构简单引入到本文的第2节中,作为本文给出讨论的理论依据与预备知识。

2 P-集合与它的结构特征^[1,2]

给定普通集合 $X = \{x_1, x_2, \dots, x_m\} \subset U$, $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_k\} \subset V$ 是 X 的属性集合,称 X^F 是 X 生成的内 P-集合(internal packet sets),简称 X^F 是内 P-集合,而且

$$X^F = X - X^- \quad (1)$$

X^- 称作 X 的 $\bar{F}$ -元素删除集合,而且

$$X^- = \{x | x \in X, \bar{f}(x) = u \in X, \bar{f} \in \bar{F}\} \quad (2)$$

到稿日期:2010-03-09 返修日期:2010-06-29 本文受山东省自然科学基金资助项目(Y2007H02)资助。

谢维奇(1973—),男,硕士,讲师,主要研究方向为数据挖掘、信息系统理论及应用, E-mail: zmdvich@yahoo.com.cn; 于秀清(1968—),女,硕士,副教授,主要研究方向为信息系统理论及应用。

如果 X^F 的属性集合 a^F 满足

$$a^F = a \cup \{a' \mid f(\beta) = a' \in a, f \in F\} \quad (3)$$

式中, $\beta \in V, \beta \in \alpha; f \in F$ 把 β 变成 $f(\beta) = a' \in \alpha; X^F \neq \emptyset$ 。

给定普通集合 $X = \{x_1, x_2, \dots, x_m\} \subset U, \alpha = \{\alpha_1, \alpha_2, \dots, \alpha_k\} \subset V$ 是 X 的属性集合, 称 X^F 是 X 生成的外 P-集合 (outer packet sets), 简称 X^F 是外 P-集合, 而且

$$X^F = X \cup X^+ \quad (4)$$

X^+ 称作 X 的 F -元素补充集合, 而且

$$X^+ = \{x' \mid u \in U, u \in X, f(u) = x' \in X, f \in F\} \quad (5)$$

如果 X^F 的属性集合 a^F 满足

$$a^F = a - \{\beta \mid \bar{f}(\alpha_i) = \beta \in \alpha, \bar{f} \in F\} \quad (6)$$

式中, $\alpha_i \in \alpha; \bar{f} \in \bar{F}$ 把 α_i 变成 $\bar{f}(\alpha_i) = \beta \in \alpha; a^F \neq \emptyset$ 。

由内 P-集合 X^F 、外 P-集合 X^F 构成的集合对, 称作普通集合 X 生成的 P-集合 (包集合, packet sets), 而且

$$(X^F, X^F) \quad (7)$$

简称 (X^F, X^F) 是 P-集合; 普通集合 X 称作 (X^F, X^F) 的基集 (基础集 ground set)。

由式(1)~式(7), 容易得到:

定理 1 P-集合 (X^F, X^F) 与它的基集合 (普通集合) X 满足

$$(X^F, X^F)_{F=F} = X \quad (8)$$

定理 2 P-集合 $\{(X_i^F, X_j^F) \mid i \in I, j \in J\}$ 与它的基集合 (普通集合) X 满足

$$\{(X_i^F, X_j^F) \mid i \in I, j \in J\}_{F=F} = X \quad (9)$$

式中, I, J 是指标集 (index set)。式(9)是 P-集合的集合对族的表示形式。

定理 3 P-集合 (X^F, X^F) 与它的基集合 (普通集合) X 满足

$$X^F \subseteq X \subseteq X^F \quad (10)$$

内 P-集合 X^F 给出一个事实: 若 X_i^F 是 X 的内 P-集合, 而且

$$X_n^F \subseteq X_{n-1}^F \subseteq \dots \subseteq X_2^F \subseteq X_1^F \quad (11)$$

则 X_i^F 的属性集 a_i^F 与 X 的属性集 a 满足

$$a_1^F \subseteq a_2^F \subseteq \dots \subseteq a_{n-1}^F \subseteq a_n^F \quad (12)$$

显然, 内 P-集合 X_i^F 具有动态特性。

利用本文第 2 节中的概念, 可给出下面第 3 节。

3 数据内搜索与搜索的依赖准则

约定 本文第 2 节中的集合 X^F, X, X^F 分别记作 $(x)^F, (x), (x)^F$; 或者 $(x)^F = X^F, (x) = X, (x)^F = X^F$ 。

定义 1 称 $(x)^F$ 是 (x) 生成的一个 $\bar{F}$ -数据, 如果 $(x)^F$ 的属性集合 a^F 与 (x) 的属性集合 a 满足

$$a^F = a \cup \{a' \mid f(\beta) = a' \in a, f \in F\} \quad (13)$$

式中, $(x)^F = \{x_1, x_2, \dots, x_r\}, (x) = \{x_1, x_2, \dots, x_m\}, r < m$ 。

定义 2 给定数据 (x) , a 是 (x) 的属性集合, 若对 a 给予属性补充, 而且

$$a_i^F = a \cup \{a' \mid \beta \in V, \beta \in \alpha, f(\beta) = a' \in \alpha\} \quad (14)$$

具有属性 a_i^F 的 $\bar{F}$ -数据 $(x)_i^F$ 称作 (x) 的一个内搜索数据, 简称 $\bar{F}$ -数据 $(x)_i^F$ 是 (x) 的一个内搜索。

其中, $(x)_i^F \subseteq (x), a \subseteq a_i^F, f \in F$ 。

定义 3 称 $\{(x)_i^F \mid i=1, 2, \dots, n\}$ 是 $\bar{F}$ -数据集, $(x)^*$ 是 $\bar{F}$ -

数据集的核, 如果

$$(x)^* = \bigcap_{i=1}^n (x)_i^F \quad (15)$$

式中, $(x)^*$ 对应的属性集合 $a^*, (x)^* \neq \emptyset$ 。

定义 4 称 ρ_i^F 是 $\bar{F}$ -数据 $(x)_i^F$ 的搜索系数 (search coefficient), 如果

$$\rho_i^F = \text{card}((x)_i^F) / \text{card}((x)) \quad (16)$$

式中, $\text{card} = \text{cardinal number}, \rho_i^F \in R^+$ 。

定义 5 若 $(x)_i^F, (x)_j^F$ 是 (x) 生成的 $\bar{F}$ -数据, 称 $(x)_i^F$ 单依赖于 $(x)_j^F$, 记作 $(x)_j^F \Rightarrow (x)_i^F$, 如果 $a_i^F \Rightarrow a_j^F$ 。

定义 6 若 $(x)_i^F, (x)_j^F$ 是 (x) 生成的 $\bar{F}$ -数据, 称 $(x)_i^F$ 双依赖于 $(x)_j^F$, 记作 $(x)_j^F \Leftrightarrow (x)_i^F$, 如果 $a_i^F \Leftrightarrow a_j^F$ 。

其中, “ $\Rightarrow$ ”, “ $\Leftrightarrow$ ”取自数理逻辑; “ $\Rightarrow$ ”与“ $\subseteq$ ”等价; “ $\Leftrightarrow$ ”与“ $=$ ”等价。

定义 7 若 $(x)_j^F \Rightarrow (x)_i^F$, 称 λ_{ij}^F 是 $(x)_i^F$ 对 $(x)_j^F$ 的依赖度, 而且

$$\lambda_{ij}^F = \text{card}((x)_j^F) / \text{card}((x)_i^F) \quad (17)$$

由定义 1~定义 7 得到:

命题 1 (x) 生成的 $\bar{F}$ -数据 $(x)_i^F, (x)_j^F$, 满足 $(x)_j^F \Rightarrow (x)_i^F$, 必有 $(x)_i^F$ 的属性集合 a_i^F 与 $(x)_j^F$ 的属性集合 a_j^F 满足 $a_i^F \Rightarrow a_j^F$; 反之亦真, $i \neq j$ 。

命题 2 (x) 生成的 $\bar{F}$ -数据 $(x)_i^F, (x)_j^F$, 满足 $(x)_j^F \Leftrightarrow (x)_i^F$, 必有 $(x)_i^F$ 的属性集合 a_i^F 与 $(x)_j^F$ 的属性集合 a_j^F 满足 $a_i^F \Leftrightarrow a_j^F$; 反之亦真, $i \neq j$ 。

由定义 1~定义 7、命题 1、命题 2 可以得到:

定理 4 ($\bar{F}$ -数据顺序内搜索定理) 若 a_i^F 是 $\bar{F}$ -数据 $(x)_i^F$ 的属性集, $i=1, 2, \dots, n$, 满足

$$a_1^F \subseteq a_2^F \subseteq \dots \subseteq a_{n-1}^F \subseteq a_n^F \quad (18)$$

则 $\bar{F}$ -数据 $(x)_i^F$ 依 $1, 2, \dots, n$ 的顺序从 (x) 中内搜索, 而且

$$(x)_n^F \subseteq (x)_{n-1}^F \subseteq \dots \subseteq (x)_2^F \subseteq (x)_1^F \quad (19)$$

证明: 若对数据 (x) 对应的属性集合 a 给予补充属性, 而且 $a_1^F = a \cup \{a_1' \mid \beta_1 \in V, \beta_1 \in \alpha, f(\beta_1) = a_1' \in \alpha, f \in F\}$, 则 $(x)_1^F$ 是 (x) 的内搜索。又 $a_1^F \subseteq a_2^F, a_2^F = a_1^F \cup \{a_2' \mid \beta_2 \in V, \beta_2 \in \alpha, f(\beta_2) = a_2' \in \alpha, f \in F\}$, $(x)_2^F$ 是 (x) 的内搜索, 且 $(x)_2^F \subseteq (x)_1^F$ 。同理可知, $\bar{F}$ -数据 $(x)_i^F$ 依 $1, 2, \dots, n$ 的顺序从 (x) 中内搜索, 而且 $(x)_n^F \subseteq (x)_{n-1}^F \subseteq \dots \subseteq (x)_2^F \subseteq (x)_1^F$ 。

定理 5 (内搜索数据辨识定理) 若 $(x)_i^F, (x)_j^F$ 是 (x) 的 $\bar{F}$ -数据, a_i^F, a_j^F 分别是 $(x)_i^F, (x)_j^F$ 的属性集, 若 $a_i^F \neq a_j^F$, 则

$$\text{IDE}_{a_i^F \neq a_j^F}((x)_i^F, (x)_j^F) \quad (20)$$

式中, $\text{IDE} = \text{identification}, (x)_i^F \neq \emptyset, (x)_j^F \neq \emptyset$ 。

证明: 若 $a_i^F \neq a_j^F$, 这里假设 $a_i^F \subseteq a_j^F, (x)_j^F \subseteq (x)_i^F$, 则 $\text{IDE}((x)_i^F, (x)_j^F) = 0$ 。

定理 6 (内搜索数据不可辨识定理) 若 $(x)_i^F, (x)_j^F$ 是 (x) 的 $\bar{F}$ -数据, a_i^F, a_j^F 分别是 $(x)_i^F, (x)_j^F$ 的属性集, 则

$$\text{UNI}_{a_i^F = a_j^F}((x)_i^F, (x)_j^F) \quad (21)$$

式中, $\text{UNI} = \text{unidentification}, (x)_i^F \neq \emptyset, (x)_j^F \neq \emptyset$ 。

证明过程与定理 5 相似, 略。

定理 7 (内搜索数据依赖属性定理) 设 $(x)_i^F$ 单依赖于 $(x)_j^F$, 即 $(x)_j^F \Rightarrow (x)_i^F$, 则它们的属性集合满足

$$a_i^F \cup \{a' \mid \beta \in V, \beta \in a_i^F, f(\beta) = a' \in \alpha, f \in F\} = a_j^F \quad (22)$$

式中, a_i^F, a_j^F 分别是 $(x)_i^F, (x)_j^F$ 的属性集合。

证明: 若 $(x)_j^F \Rightarrow (x)_i^F$, 则 $a_j^F \subseteq a_i^F$, 即存在 $\{a' \mid \beta \in V, \beta \in a_i^F, f(\beta) = a' \in \alpha, f \in F\}$, 使得 $a_i^F \cup \{a' \mid \beta \in V, \beta \in a_i^F, f(\beta) = a' \in \alpha, f \in F\} = a_j^F$ 。

定理 8 (内搜索数据依赖传递定理) 若 $(x)_i^F, (x)_j^F, (x)_k^F$ 是 (x) 内搜索的 $\bar{F}$ 数据, 且 $(x)_k^F \Rightarrow (x)_j^F, (x)_j^F \Rightarrow (x)_i^F$, 则 $(x)_k^F \Rightarrow (x)_i^F$ (23)

证明: $(x)_i^F, (x)_j^F, (x)_k^F$ 对应的属性集分别是 a_i^F, a_j^F, a_k^F 。由 $(x)_k^F \Rightarrow (x)_j^F$, 知 $a_k^F \subseteq a_j^F$, 由 $(x)_j^F \Rightarrow (x)_i^F$, 知 $a_j^F \subseteq a_i^F, a_k^F \subseteq a_i^F$, $(x)_k^F \Rightarrow (x)_i^F$ 。

定理 9 $(\{(x)_i^F\}, \Rightarrow)$ 是序关系, 或者

$$(x)_n^F \Rightarrow \dots \Rightarrow (x)_1^F \Rightarrow (x)_0^F \quad (24)$$

式中, 记 $(x)_0^F = (x), i=0, 1, \dots, n$ 。

证明: $\Rightarrow$ 在 $\{(x)_i^F\}$ 中具有自反性、反对称性、传递性, $(\{(x)_i^F\}, \Rightarrow)$ 是偏序关系。又 $\forall (x)_i^F, (x)_j^F \in \{(x)_i^F\}$ 是可比的, $(\{(x)_i^F\}, \Rightarrow)$ 是序关系。

推论 1 $(x)_n^F = (x)^*$ 是 $(\{(x)_i^F\}, \Rightarrow)$ 的最小元, $(x)_0^F = (x)$ 是 $(\{(x)_i^F\}, \Rightarrow)$ 的最大元。

定理 10 $(\{a_i^F\}, \Rightarrow)$ 是序关系, 或者

$$a_0^F \Rightarrow a_1^F \Rightarrow \dots \Rightarrow a_n^F \quad (25)$$

式中, 记 $a_0^F = \alpha, i=0, 1, \dots, n$ 。

证明与定理 9 相似, 略。

推论 2 a_0^F 是 $(\{a_i^F\}, \Rightarrow)$ 的最小元, $a_n^F = \alpha^*$ 是 $(\{a_i^F\}, \Rightarrow)$ 的最大元。

定理 11 (搜索系数链定理) $(x)_i^F$ 是 (x) 生成的 $\bar{F}$ 数据, ρ_i^F 是 $(x)_i^F$ 的搜索系数, 如果 $(x)_n^F \subseteq (x)_{n-1}^F \subseteq \dots \subseteq (x)_2^F \subseteq (x)_1^F$, 则

$$\rho_n^F \leq \rho_{n-1}^F \leq \dots \leq \rho_1^F \quad (26)$$

证明: 若 $(x)_n^F \subseteq (x)_{n-1}^F \subseteq \dots \subseteq (x)_2^F \subseteq (x)_1^F$, $\text{card}((x)_n^F) \leq \text{card}((x)_{n-1}^F) \leq \dots \leq \text{card}((x)_1^F)$, $\text{card}((x)_n^F)/\text{card}((x)) \leq \dots \leq \text{card}((x)_1^F)/\text{card}((x))$, $\rho_n^F \leq \rho_{n-1}^F \leq \dots \leq \rho_1^F$ 。

定理 12 (单位离散区间内点定理) 若 $(x)_i^F$ 是 (x) 的 $\bar{F}$ 数据, ρ_i^F 是 $\bar{F}$ 数据 $(x)_i^F$ 的搜索系数, 则 ρ_i^F 是单位离散区间 $(0, 1)$ 的一个内点, 而且

$$\rho_i^F \in (0, 1) \quad (27)$$

证明: 设 $\rho = 1$ 是 (x) 的搜索系数, $0 < \rho_i^F = \text{card}((x)_i^F)/\text{card}((x)) < 1$, ρ_i^F 是单位离散区间 $(0, 1)$ 的一个内点, $\rho_i^F \in (0, 1)$ 。

定理 13 (搜索系数关系定理) 若 $(x)_i^F, (x)_j^F, (x)_k^F$ 是 (x) 内搜索的 $\bar{F}$ 数据, 它们的搜索系数满足 $\rho_k^F \leq \rho_j^F \leq \rho_i^F$, 则

$$(x)_k^F \Rightarrow (x)_j^F \Rightarrow (x)_i^F \quad (28)$$

证明: 若 $\rho_k^F \leq \rho_j^F \leq \rho_i^F$, $(x)_k^F \subseteq (x)_j^F \subseteq (x)_i^F, a_k^F \subseteq a_j^F \subseteq a_i^F$, $(x)_k^F \Rightarrow (x)_j^F \Rightarrow (x)_i^F$ 。

定理 14 (最小搜索系数存在定理) 在 $\bar{F}$ 数据 $(x)_i^F$ 中, 一定存在数据 $(x)_k^F$ 的搜索系数最小, 或者

$$\rho_k^F = \min_{i=1}^n (\rho_i^F) \quad (29)$$

证明: 若 $(x)_n^F \subseteq (x)_{n-1}^F \subseteq \dots \subseteq (x)_2^F \subseteq (x)_1^F, \rho_n^F \leq \dots \leq \rho_1^F$,

$$(x)^* = \bigcap_{i=1}^n (x)_i^F, \rho_k^F = \min_{i=1}^n (\rho_i^F) = \rho(x)^*。$$

由定义 1—定义 7、命题 1、命题 2、定理 4—定理 14 得到如下算法。

$\bar{F}$ -数据内搜索迭代算法

Step 1 数据 $(x) = \{x_1, x_2, \dots, x_m\}$, 对应的属性集合 $\alpha = \{a_1, a_2, \dots, a_k\}, i=0$;

Step 2 α 中补充属性 $\beta_1, a_1^F = \alpha \cup \beta_1, (x)_1^F = (x) - \Delta(x)_1, i=1$;

Step 3 a_i^F 中补充属性 $\beta_2, a_2^F = a_i^F \cup \beta_2, (x)_2^F = (x)_1^F - \Delta(x)_2, i=2$;

Step 4 if $\rho_i^F > \theta, a_i^F = a_{i-1}^F \cup \beta_i, (x)_i^F = (x)_{i-1}^F - \Delta(x)_i, i=i+1$ 。否则, 算法停止。

其中, $\Delta(x)_i$ 是内搜索余量; θ 是内搜索阈值; 取值事先设定, $0 < \theta < 1$ 。

$\bar{F}$ -数据的依赖-内搜索准则

如果数据 (x) 的属性集合 α 之外的属性集合 $\beta_i \subseteq \beta_j \subseteq \beta_k$ 被依次补充到 α 内, 则 (x) 的 $\bar{F}$ -数据 $(x)_i^F, (x)_j^F, (x)_k^F$ 依 $(x)_i^F, (x)_j^F, (x)_k^F$ 的顺序依次被搜索, 而且

$$(x)_k^F \Rightarrow (x)_j^F \Rightarrow (x)_i^F \quad (30)$$

$\forall i \in (i, j, k)$

$$(x)_i^F \cup (x) = (x) \quad (31)$$

4 数据内搜索在信息搜索中的应用

为了简单, 又不失一般性, 本节给出数据内搜索在信息搜索中的应用。以 Baidu 上搜索为例, 依次输入查询关键词 a_1, a_2, a_3, a_4 , 得到数据表 1。

表 1 内搜索数据表

α	X	card(X)
$\alpha = \{a_1\}$	(x)	100,000,000
$\alpha_1^F = \{a_1, a_2\}$	$(x)_1^F$	12,200,000
$\alpha_2^F = \{a_1, a_2, a_3\}$	$(x)_2^F$	953,000
$\alpha_3^F = \{a_1, a_2, a_3, a_4\}$	$(x)_3^F$	8,400

其中, 取 $\theta = 10^{-4}$, $(x)_3^F \subseteq (x)_2^F \subseteq (x)_1^F \subseteq (x), \alpha \subseteq \alpha_1^F \subseteq \alpha_2^F \subseteq \alpha_3^F$, 依 $(x), (x)_1^F, (x)_2^F, (x)_3^F$ 顺序依次被搜索。

$$\rho_1^F = \text{card}((x)_1^F)/\text{card}((x)) = 0.122$$

$$\rho_2^F = \text{card}((x)_2^F)/\text{card}((x)) = 0.000953$$

$$\rho_3^F = \text{card}((x)_3^F)/\text{card}((x)) = 0.000084 < \theta$$

$(x)_2^F$ 就是要得到的搜索结果。 $(x)_1^F$ 对 $(x)_2^F$ 的依赖度是

$$\lambda_{12}^F = \text{card}((x)_2^F)/\text{card}((x)_1^F) = 0.0781$$

对例子的讨论:

1) 本节的例子是在搜索引擎上实际搜索得到的结果。 a_1, a_2, a_3, a_4 是搜索关键词, 这里作为属性, $\alpha_1 - \alpha_4$ 的名称略。

2) (x) 是搜索关键词 a_1 得到的结果, 作为基集合。 $(x)_1^F, (x)_2^F, (x)_3^F$ 是依次增加关键词 a_2, a_3, a_4 后得到的结果。

3) 内搜索阈值 θ 取 10^{-4} , $(x)_2^F$ 是在 $\theta = 10^{-4}$ 条件下得到的搜索结果。

结束语 本文利用 P-集合的动态特性研究数据内搜索的过程和特征, 给出数据内搜索的依赖关系及度量。P-集合普遍存在于动态数据系统中, 它为研究动态数据系统提供了

参 考 文 献

[1] Shi Kaiquan. P-sets and its applications[J]. An International Advances in Systems Science and Applications,2009,9(2):209-219

[2] 史开泉. P-集合[J]. 山东大学学报:理学版,2008,43(11):77-84

[3] 史开泉,张丽. 内 P-集合与数据外-恢复[J]. 山东大学学报:理学版,2009,44(4):8-14

[4] 汤积华,陈保会,史开泉. P-集合与(F,F)-数据生成-辨识[J]. 山东大学学报:理学版,2009,44(9):50-55

[5] 于秀清. P-集合的识别与筛选[J]. 山东大学学报:理学版,2010,45(1):94-98

[6] 张飞,陈萍,张丽. P-集合的 P-分离与应用[J]. 山东大学学报:理

学版,2010,45(3):18-22

[7] Shi Kaiquan,LI Xiuhong. Camouflaged informatiation and its on identification and its applications[J]. An International Journal Advances in System Science and Applications,2010,10(2):208-216

[8] Zhang li, Cui Yuquan. Outer P-sets and data internal-recover [J]. An International Journal Advances in System Science and Applications,2010,10(2):229-236

[9] Lin Hong-kang, Li Yu-ying. P-sets and its P-separation theorems[J]. An International Journal Advances in System Science and Applications,2010,10(2):245-251

[10] Zhang Guan-yu,LI En-zhong. Information gene and its information knock-out/knock-in[J]. An International Journal Advances in System Science and Applications,2010,10(2):267-27

(上接第 231 页)

验,去掉其中非完备的对象,该数据库共有 5644 个对象,记录蘑菇的有 22 个条件属性、1 个决策属性。将 mushroom 数据库看作决策表,选择 4644 个对象作为基准决策表,从剩下的 1000 个对象中依次增加 100,200,300,400,500,600,700,800,900,1000 个对象作为增量,实验结果如表 2 所列。

表 2 非增量与增量约简算法的实验结果对比表

对象数	ARC		IARC	
	约简结果	时间(S)	约简结果	时间(S)
4744	{2,3,4,6,8,10,13,14,15,16,21,22}	7774.39	{2,3,4,6,8,10,13,14,15,16,21,22}	5564.67
4844	{2,3,4,6,8,10,13,14,15,16,21,22}	8035.61	{2,3,4,6,8,10,13,14,15,16,21,22}	5827.67
4944	{2,3,4,6,8,10,13,14,15,16,21,22}	8459.110	{2,3,4,6,8,10,13,14,15,16,21,22}	6151.89
5044	{2,3,4,6,8,10,13,14,15,16,21,22}	9311.061	{2,3,4,6,8,10,13,14,15,16,21,22}	6447.88
5144	{2,3,4,6,8,10,13,14,15,16,21,22}	9659.641	{2,3,4,6,8,10,13,14,15,16,21,22}	6772.91
5244	{2,3,4,6,8,10,13,14,15,16,21,22}	10279.69	{2,3,4,6,8,10,13,14,15,16,21,22}	7123.34
5344	{2,3,4,6,8,10,13,14,15,16,21,22}	10750.16	{2,3,4,6,8,10,13,14,15,16,21,22}	7507.74
5444	{2,3,4,6,8,10,13,14,15,16,21,22}	11527.95	{2,3,4,6,8,10,13,14,15,16,21,22}	7865.13
5544	{2,3,4,6,8,10,13,14,15,16,21,22}	12480.84	{2,3,4,6,8,10,13,14,15,16,21,22}	8340.89
5644	{2,3,4,6,7,8,10,13,14,15,16,21,22}	12908.11	{2,3,4,6,7,8,10,13,14,15,16,21,22}	8829.38

从图 1 可见,本文的算法优于传统的非增量算法,原因正是本文的算法可以有效利用已经计算过的条件熵和约简结果,通过更新已有的约简结果来计算新约简。

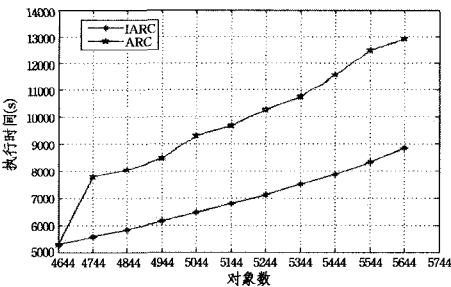


图 1 算法执行时间

结束语 本文通过分析决策表对象动态增加时条件熵关

于条件属性的变化机制,提出了一种基于条件熵的属性约简增量式更新算法。算法充分利用了原决策表的约简属性和条件熵的信息,明显地减少了增量更新属性约简的计算量,为动态更新决策表的属性约简提供了一种快速有效的新途径。

参 考 文 献

[1] Pawlak Z. Rough sets theoretical aspects of reasoning about data [M]. Kluwer Academic Publisher,1991

[2] 张文修,吴伟业,梁吉业,等. 粗糙集的理论与方法[M]. 北京:科学出版社,2001

[3] 梁吉业,李德玉. 信息系统中的不确定性与知识获取[M]. 北京:科学出版社,2005

[4] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domain for neural networks[J]. Computational Intelligence,1995,11(2):339-347

[5] 刘少辉,盛秋戡,吴斌,等. Rough 集高效算法的研究[J]. 计算机学报,2003,26(5):524-529

[6] Guan J W, Bell D A. Rough computational methods for information systems[J]. Artificial Intelligences,1998,105(1/2):77-103

[7] Woblewski J. Finding minimal reducts using genetic algorithm [R]. Warsaw University of Technology; ICS Research Report, 16/95,1995

[8] 林俊伟,叶东毅. 基于领域辨识矩阵的属性约简增量式算法[J]. 计算机应用,2009,29(11):119-121

[9] Hu F, Wang G Y, Huang H, et al. Incremental attribute reduction based on elementary sets[C]//Processing of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Regina, Canada,2005:183-193

[10] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报,2007,30(5):815-823

[11] 梁吉业,魏巍,钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践,2008,28(4):81-89

[12] Liang J Y, Xu Z B. The algorithm on knowledge reduction in incomplete information system[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based System,2002,10(1):95-103

[13] 梁吉业,钱宇华. 信息系统中的信息粒与熵理论[J]. 中国科学(E 辑),2008,38(12):2048-2065

[14] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766