TẠP CHÍ KHOA HỌC TÍNH TOÁN VÀ ĐIỀU KHIỂN Tập I (1985) SỐ 3 (25–28)

PHƯƠNG PHÁP MCA TRONG LƯU TRỮ VÀ TÌM TIN

NGUYỄN CHÍ THÀNH

MỞ ĐẦU

Trong quá trình dùng máy tính điện tử để xử lý thông iin đã xuất hiện một số phương pháp tìm kiếm thông dụng như tìm kiếm liên tiếp, tuần chí, các phương pháp biến đổi địa chỉ và truy nhập trực tiếp, phương pháp tệp ngược, hàm băm, phương pháp Salton, Lum, v.v... Mỗi phương pháp đều có những ưu, nhược điểm riêng và được vận dụng trong từng quá trình xử lý cụ thể. Mới đây, do nhu cầu đối thoại giữa người và máy đã xuất hiện phương pháp MCA (phương pháp chỉ tiêu ghép). Cũng do mới hình thành nên phương pháp MCA còn đang được hoàn thiện về lý luận và tiếp tục khảo nghiệm trong thực tiễn. Trong khuôn khổ bài báo này, tác giả muốn làm rõ thêm một số vấn đề về lý luận của phương pháp MCA như chứng minh tính duy nhất của địa chỉ đối tượng trong tệp lưu trữ, tính duy nhất của giá trị thuộc tính khi cần khôi phục lại chúng.

PHƯƠNG PHÁP MCA

Cho hệ tin:

$$S = \langle X, A, V, f \rangle$$

Trong đó

X – tập các đối tượng hay tập các tài liệu;

A – tập các thuộc tính của đối tượng

$A = a_1, a_2, ..., a_n, n \geqslant 1$;

V = tập các mô tả hay tập các giá trị của thuộc tính

$V = \bigcup_{a \in A} V_a$

trong đó V_a là miền giá trị của thuộc tính $a \in A$; f – hẩm thông tin ánh xạ từ $X \times A$ lên V sao cho

$$f(x, a) \in V_a \quad \forall x \in X \text{ và } \forall a \in A.$$

Với mỗi $x \in X$, hàm f_x là một ánh xạ từ A lên V sao cho $f_x(a) = f(x, a)$. Hàm f_x được xem như là thông tin về đối tượng x trong hệ tin S.

Đặt V_i là tập các mô tả của thuộc tính a_i;

Trong mỗi tập V_i, các mô tả đều được đánh số: 0, 1, 2, 3, ..., $p_i - 1$.

Gọi P_i là lực lượng của V_i.

Thông tin về một đối tượng được biểu diễn bằng bộ

$$b = \langle b_1, b_2, ..., b_n \rangle, \quad b_i \in V_i.$$

Vị trí của đối tượng x trong tập tin được xác định bởi số.

$$d(x) = \sum_{i=1}^{n} \overline{b_i} u_i \qquad (1)$$

Số $d(x)$ còn gọi là địa chỉ của đối tượng x trong hệ tin.

$\overline{b_i}$ là số thứ tự của giá trị b_i trong tập V_i.

Với mỗi đối tượng x ta đặt $D(x) = \langle \overline{b_1}, \overline{b_2}, \ldots, \overline{b_n} \rangle$ là bộ đặc trưng cho x

$$\left.\begin{aligned} u_1 &= P_2 \times P_3 \times \ldots P_n \\ u_2 &= P_3 \times \ldots P_n \\ &\ldots\ldots\ldots\ldots \\ u_{n-1} &= P_n \\ u_n &= 1 \end{aligned}\right\} \qquad (2)$$

Từ số $d(x)$ của công thức (1) ta có thể xác định một hay toàn bộ các giá trị $\overline{b_i}$ theo phương pháp sau đây:

$$\left.\begin{aligned} \overline{b_1} &= \text{entier}\ \frac{d(x)}{u_1} \\ \overline{b_2} &= \text{entier}\ \frac{d(x) - \overline{b_1} u_1}{u_2} \\ &\ldots\ldots\ldots\ldots \\ \overline{b_i} &= \text{entier}\ \frac{d(x) - \sum_{j=1}^{i-1} \overline{b_j} u_j}{u_i} \\ &\ldots\ldots\ldots\ldots \\ \overline{b_n} &= \text{entier}\ \frac{d(x) - \sum_{j=1}^{n-1} \overline{b_j} u_j}{u_n} \end{aligned}\right\} \qquad (3)$$

Ta sẽ chứng minh rằng các $\overline{b_i}$ tính được là duy nhất.

Thật vậy, từ (2) ta có:

$$P_j u_j = u_{j-1} \qquad (2')$$

Mặt khác ta có

$$\overline{b_j} \leqslant P_j - 1 \qquad (3')$$

và

$$d(x) = \overline{b_1} u_1 + \overline{b_2} u_2 + \ldots + \overline{b_n} u_n.$$

Giả sử với một i nào đó ($1 \leqslant i < n$) ta đã tính được $\overline{b_1}, \overline{b_2}, \ldots, \overline{b_{i-1}}$, khi đó

$$\frac{d(x) - \sum_{j=1}^{i-1} \overline{b_j} u_j}{u_i} = \overline{b_i} + \frac{\sum_{j=i+1}^{n} \overline{b_j} u_j}{u_i} \qquad (4)$$

Ta sẽ chứng tỏ rằng:

$$\frac{\sum_{j=i+1}^{n} \overline{b_j} u_j}{u_i} < 1.$$

Theo (3') ta có thể đánh giá đại lượng trên như sau:

$$\frac{\sum_{j=i+1}^{n} \overline{b_j} u_j}{u_i} \leqslant \frac{\sum_{j=i+1}^{n} (P_j - 1) u_j}{u_i} =$$

$$= \frac{\sum_{j=i+1}^{n} P_j u_j - \sum_{j=i+1}^{n} u_j}{u_i} = \frac{\sum_{j=i+1}^{n} u_{j-1} - \sum_{j=i+1}^{n} u_j}{u_i} =$$

$$= \frac{u_i - u_n}{u_i} = 1 - \frac{1}{u_i} < 1.$$

Trường hợp i = n cho ta

$$\text{entier}\ \frac{d(x) - \sum_{j=1}^{n-1} \overline{b_j}\, u_j}{u_n} =$$

$$= \text{entier}\ \frac{\sum_{j=1}^{n} \overline{b_j}\, u_j - \sum_{j=1}^{n-1} \overline{b_j} u_j}{u_n} = \text{entier}\ \frac{\overline{b_n}\, u_n}{u_n} = \overline{b_n}.$$

Tính duy nhất của các $\overline{b_i}$ theo (3) đã được chứng minh.

Định nghĩa. Với mỗi hệ tin $S = \langle X, A, V, f \rangle$, ta xây dựng quan hệ «đứng trước x» trong tập X như sau:

Giả sử với $x, y \in X$ ta có:

$$D(x) = \langle \overline{b_1}, \overline{b_2}, \ldots, \overline{b_n} \rangle$$
$$D(y) = \langle \overline{c_1}, \overline{c_2}, \ldots, \overline{c_n} \rangle$$

Ta có x đứng trước y nếu $\exists\, i : 1 < i \leqslant n$

thỏa mãn 2 điều kiện sau:

(1) $\overline{b_j} = \overline{c_j}$ với $j < i$

(2) $\overline{b_i} < \overline{c_i}$

Ví dụ:

$D(x) = \langle 1, 2, 6, 80, 800 \rangle$

$D(y) = \langle 1, 2, 8, 2, 10 \rangle$

ta có x đứng trước y vì $6 < 8$.

Định lý. x đứng trước y khi và chỉ khi $d(x) < d(y)$

Chứng minh:

Theo công thức (1)

$$d(x) = \overline{b_1}u_1 + \overline{b_2}u_2 + \ldots + \overline{b_n}u_n \qquad (5)$$

$$d(y) = \overline{c_1}u_1 + \overline{c_2}u_2 + \ldots + \overline{c_n}u_m \qquad (6)$$

trong đó $\overline{b_i}, \overline{c_i} \in \{0, 1, 2, \ldots, P_i - 1\}$; $1 \leqslant i \leqslant n$.

Giả sử với i nào đó ta có: $\overline{b_i} < \overline{c_i}$,

$\overline{b_j} = \overline{c_j}$ với $j < i$. Ta sẽ chứng minh rằng $d(x) < d(y)$.

Không giảm ý nghĩa tổng quát, ta có thể giả thiết i = 1, khi đó từ bất đẳng thức $\overline{b_1} < \overline{c_1}$ ta suy ra

$$\overline{c_1} \geqslant \overline{b_1} + 1 \qquad (7)$$

Từ (6) và (7) ta có

$$d(y) \geqslant (\overline{b_1} + 1)\, u_1 \qquad (8)$$

Mặt khác từ (5) và (3') ta có

$$d(x) \leqslant \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + (P_n - 1)\, u_n \qquad (9)$$

Ta ký hiệu vế phải của (8) là M, vế phải của (9) là N, đồng thời thay $u_n = 1$ và $P_n - 1$ bằng P_n trong (9) ta được

$$\begin{aligned}
N &< \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + (P_{n-1} - 1)\, u_{n-1} + P_n = \\
&= \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + (P_{n-1} - 1)\, P_n + P_n = \\
&= \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + (P_{n-2} - 1)\, u_{n-2} + P_{n-1}P_n = \\
&= \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + (P_{n-2} - 1)\, P_{n-1}P_n + P_{n-1}P_n = \\
&= \overline{b_1}u_1 + (P_2 - 1)\, u_2 + \ldots + P_{n-2}P_{n-1}P_n = \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&= \overline{b_1}u_1 + P_2P_3 \ldots P_n = (\overline{b_1} + 1)u_1 = M.
\end{aligned}$$

Như vậy là $d(y) \geqslant M > N \geqslant d(x)$

tức là $d(x) < d(y)$.

Ngược lại, nếu $d(x) < d(y)$ ắt phải tồn tại số nguyên i để cho

$$(\overline{b_1} = \overline{c_1}) \wedge (\overline{b_2} = \overline{c_2}) \wedge \ldots \wedge (\overline{b_{i-1}} = \overline{c_{i-1}}) \wedge (\overline{b_i} = \overline{c_i}).$$

Nếu không tồn tại số nguyên i như vậy thì phải có số nguyên j mà

$$(\overline{b_1} = \overline{c_1}) \wedge (\overline{b_2} = \overline{c_2}) \wedge \ldots \wedge (\overline{b_{j-1}} = \overline{c_{j-1}}) \wedge (\overline{b_j} > \overline{c_j}),$$

nhưng khi đó $d(x) > d(y)$ trái với giả định ban đầu.

Hệ quả. Nếu x khác y thì $d(x)$ khác d(y).

Những điều khẳng định trên đây cho phép ta thực hiện địa chỉ hóa các đối tượng của hệ tin: ứng với mỗi bộ mô tả cụ thể chỉ tồn tại duy nhất một địa chỉ lưu trữ, hay nói khác là trong hệ tin, mỗi vị trí chỉ xác định duy nhất một đối tượng cụ thể.

Định lý trên còn cho phép ta mở rộng khả năng sử dụng phương pháp MCA, giải ải quyết các câu hỏi ở dạng bất đẳng thức, vì rằng trật tự từ vựng của các bộ mô tả được thể hiện bằng trật tự của bảng địa chỉ, do đó tập các bản ghi thỏa mãn một câu hỏi cụ thể sẽ được phân bố liên tiếp nhau.

Tác giả xin chân thành cảm ơn giáo sư cấp 1 Hồ Thuần và PTS Nguyễn Xuân Huy (Viện Khoa học tính toán và Điều khiển) đã đọc và giúp đỡ chỉnh lý bài này.

Nhận ngày 17-8-1984

TÀI LIỆU THAM KHẢO

1. **Zdzislaw Pawlak**, Information Systems ICS PAS Reports, N°338, Warszawa 1978.

2. **Zdzislaw Pawlak**, Distributed Information Systems ICS PAS Reports, N°370, Warszawa 1979.

ABSTRACT

The paper presents one proof of the non – repetition of address for object in MCA–method and the unicity of attributed value, which is obtained from the deloppment of addresses.