VẤN ĐỀ TỐI THIỂU HÓA HỆ ĐIỀU CHỈNH THÔNG TIN

ĐỖ ĐỨC GIÁO, VŨ NGỌC LOÃN

Trong [7] tác giả đã đưa ra định nghĩa tập các cây thông tin TREE. Với mỗ cây $T \in TREE$, ta lập bộ ba $S(T) = [R, D_T, \delta_T]$ và gọi là hệ điều chỉnh thông tin ở đây R — tập các công thức lôgic, D_T — tập tất cả các thông tin khác nhau từng cặp có mặt trong cây T, còn δ_T là ánh xạ từ tập R vào tập 2^{D_T}. Như vậy tập các hệ điều chỉnh thông tin $S(TREE) = [R, D_T, \delta_T]/R \subseteq FORM, T \in TREE]$ là một lớp các retrieval Systems mà SALTON đã định nghĩa trong [4]. Trong bài này, tác sẽ nghiên cứu vấn đề tối thiểu hóa cấu trúc hệ điều chỉnh thông tin ứng với một cây thông tin cho trước.

1. Cây thông tin với các tập công thức lôgic

Z — tập các số nguyên. D — tập các Documents. $D^+ =: DU[\tau]$ τ — kí hiệu cây rỗng. Tập các cây thông tin định nghĩa trong [7] kí hiệu là TREE.

RESULT : TREE x Z → D^+ định nghĩa như sau:

1. RESULT (d, l) = d, $\forall d \in D^+$ và $\forall l \in Z$.

2 RESULT $(K < T_1, T_2, T_3 >, l) \underset{đn}{=}$ $\begin{cases} \text{RESULT}(T_1, l) \text{ nếu } K < 1 \\ \text{RESULT}(T_2, l) \text{ nếu } K = 1 \\ \text{RESULT}(T_3, l) \text{ nếu } K > 1 \end{cases}$

Hai cây T_1 và T_2 là tương đương nhau $(T_1 \simeq T_2) \underset{đn}{\Leftrightarrow} \forall l \in Z$

ta có $\text{RESULT}(T_1, l) = \text{RESULT}(T_2, l)$

Giả sử $p, q \in Z$. x là đối số nhận giá trị trong Z.

INEQU = : [x ⩽ p, x > q/p, q ∈ Z].

t là kí hiệu đúng (nhận giá trị 1). f là kí hiệu sai (nhận giá trị 0)

Ta định nghĩa các công thức:

1. Mỗi phần tử $u(x) \in$ INEQU gọi là một công thức (công thức cơ bản). Kí hiệu t, f cũng là những công thức.
2. Giả sử H(x), $H_1(x)$. $H_2(x)$ là các công thức. Khi đo dãy kí hiệu

$(\neg H(x))$, $(H_1(x) \square H_2(x))$. ở đây $\square \in [\vee, \wedge, \rightarrow, \leftrightarrow]$.

Cũng là những công thức. Các kí hiệu $\neg$ (non), $\vee$ (vel), $\wedge$ (et), $\rightarrow$ (Seq) và $\rightarrow$ (Seq) và $\leftrightarrow$ (Saq) là các kí hiệu hàm logic. Định nghĩa các hàm này xem trong [1].

Kí hiệu tập tất cả các công thức lôgic định nghĩa như trên là FORM. b là ánh xạ từ tập $INEQU_x Z$ vào tập [0,1] và được định nghĩa như sau:

$$b(u(x), K) = \begin{cases} 1 \text{ nếu } u(K) \text{ đúng} \\ 0, \text{ nếu } u(K) \text{ sai}, \end{cases}$$

đây $u(x) \in INEQU$, $K \in Z$

Tất nhiên ta có thể mở rộng ánh xạ trên thành ánh xạ Val: $FORM_x Z \to [0,1]$ như sau:

1) $Val(u(x), k) = b(u(x), k)\ \forall u(x) \in INEQU, \forall k \in Z.$

$Val(f, k) = 0, \forall k \in Z.$

$Val(t, k) = 1, \forall k \in Z.$

2) $Val(\neg H(X), k) = non\ Val(H_1(x), k))$

$Val(H_1(x) \square H_2(x), k) = 0\ (Val(H_1(x), k))\ val\ H_2(x)k))\ Val.(H_2(x);\ k$

đây $(\square, 0) \in [(\vee, vel), (\wedge, et), (\to, Seq), (-), Seq), (\leftrightarrow, Saq)]$

Giả sử $H_1(x), H_2(x) \in FORM$. Ta nói $H_1(x)$ là tương đương (tương đương yếu) với $H_2(x)$, kí hiệu $H_1(x) \sim H_2(x)$ ($H_1(x) \sim H_2(x)$) khi và chỉ khi

$Val(H_1(x), k) = Val(H_1(x), k)$ $(val(H_1(x), k) \leqslant (val(H_2(x), k))\ \forall k \in Z$. Bây giờ ta định nghĩa ánh xạ $RETR: TREE \times FORM \to 2^{D^+}$ như sau:

$$RETR(T, H(x)) = [RESULT(T, K)/k \in Z,\ T \in TREE,\ Val(H(x), k) = 1].$$

Đặt $RETR(T, H(x)) =: \sigma_T(H)$. Khi đó dễ dàng thấy:

1. Nếu $H_1(x) \sim H_2(x)$ thì $\sigma_T(H_1) = \sigma_T(H_2)$
2. Nếu $H_1(x) \sim H_2(x)$ thì $\sigma_T(H_1) \subseteq \sigma_T(H_2)$
3. $\sigma_T(f) = \phi$ (tập rỗng)
4. $\sigma_T(\neg H) = \sigma_T(x) \setminus \sigma_T(H)$
5. $\sigma_T(H_1 \vee H_2) = \sigma_T(H_1) \cup \sigma_T(H_2)$
6. $\sigma_T(H_1 \wedge H) = \sigma_T(H_1) \cap \sigma_T(H_2)$

Giả sử $T, T' \in TREE$ và $R \subseteq FORM$.

Ta nói T là R — tương đương với T' ($T \underset{R}{\approx} T'$) $\Leftrightarrow \underset{đn}{\forall} H \in R$ ta có

$$RETR(T, H) = RETR(T', H).$$

Hiển nhiên nếu $T \approx T'$ thì $T \underset{R}{\approx} T'$

2. Hệ điều chỉnh thông tin và vấn đề tối thiểu hóa cấu trúc của nó.

Với mỗi $T \in TREE$ và $R \subseteq FORM$, kí hiệu D_T là tập tất cả các thông tin có mặt trong T và khác nhau từng cặp.

$\sigma_T: R \to 2^{D_T}$ ở đây $\sigma_T =: TREE \times R \to 2^{D_T}$.

Khi đó bộ ba $S(T) = [R, D_T, \sigma_T$ với R là tập các ngôn ngữ vào, D_T là tập các thông tin ra và σ_T là ánh xạ điều chỉnh thông tin) được gọi là hệ điều chỉnh thông tin ứng với cây T.

$$S(TREE) =: [\,[R, D_T, \sigma_T]/R \subseteq FORM,\ T \in TREE]$$

là tập tất cả các hệ điều chỉnh thông tin ứng với TREE.

Định nghĩa 1.

Giả sử $S(T) = [R, D_T, \sigma_T] \in S(TREE)$

1. Tập $A \subseteq D_T$ gọi là tập mô tả được trong $S(T) \overset{đn}{\Leftrightarrow} \exists H \in R$ sao cho $\sigma_T(H) = A$. Tập $D(S(T)) = [A \mid A \subseteq D_T \wedge \exists H \in R \wedge \sigma_T(H) = A]$ gọi là tập tất cả các tập mô tả được trong $S(T\ S(T))$.

2. Hệ $S(T) = [R, D_T, \sigma_T]$ gọi là hệ mô tả được $\overset{đn}{\Leftrightarrow} \forall d \in D_T\ \exists H \in R \wedge \sigma_T(H) = [d]$.

3. Phần tử $d \in D_T$ là dư thừa trong $S(T) \overset{đn}{\Leftrightarrow} \forall H \in R$ ta luôn luôn có

$$\sigma_T(H) \neq [d].$$

Định nghĩa 2.

1. Ta nói $S(T_1)$ tương đương với $S(T_2)$ $(S(T_1) \approx S(T_2)) \overset{đn}{\Leftrightarrow} T_1 \approx T_2$

2. Ta nói $S(T_1)$ là R — tương đương với $S(T_2)$ $(S(T_1) \underset{R}{\approx} S(T_2)) \overset{đn}{\Leftrightarrow} T_1 \underset{R}{\approx} T_2$

3. Ta nói $S(T_1)$ đồng nhất với $S(T_2)$ $(S(T_1) \equiv S(T_2)) \overset{đn}{\Leftrightarrow} T_1 \equiv T_2$ và $R_1 \equiv R$

Định lý 1

1. $\forall S(T) \in S(TREE)\ \exists S(T') \in S(TREE)$ sao cho

$S(T) \approx S(T')$ và $S(T'')$ là mô tả được.

2. Hệ $S(T) = [R, D_T, \sigma_T]$ là mô tả được khi và chỉ khi mỗi tập con của D_T là mô tả được trong $S(T)$.

Chứng minh

1. Để chứng minh phần này, ta đưa vào khái niệm cây chuẩn N. N gọi là cây chuẩn nếu nó có dạng là d $(d \in D^+)$ hoặc $K < d_0, d_1, d_2 >$ với $d_0 \neq d_1$, hoặc $k < d_0, d_1, k+1 < \tau, d_2, \ldots, k+s < \tau, d_{s+1}, d_{s+2} > \ldots >$ với $d_0 \neq d_1$ và $d_{s+1} \neq d_{s+2}$ $(s \geq 1)$. Đối với T trong $S(T)$ có tồn tại duy nhất cây chuẩn N sao cho $N \approx T$. Xem trong [7]. Kiểm tra lại $S(N) = [R, D_N, \sigma_N]$ là hệ mô tả được, và $S(T) \approx S(N)$

2. Dựa vào định nghĩa hệ mô tả được và sử dụng các tính chất đã đưa ra của ánh xạ σ_T trong phần I.

Định nghĩa 3

Giả sử $S(T_0) \in S(TREE)$. Hệ $S(T_0) = [R, D_{T_0}, \sigma_{T_0}]$ được gọi là D — tối giản $\Leftrightarrow dn[D_{T_0}] \leq \min [[D_T]/S(T) = [R, D_T, \sigma_T] \in S(TREE)]$

$S(T_0) \approx S(T)$

Định lý 2

1. $\forall S(T) \in S(TREE)\ \exists S(T_0)$ sao cho $S(T) \approx S(T_0)$ và $S(T) \approx S(T_*)$ và $S(T_*)$ là D — tối giản.

2. Hệ $S(T)$ là D — tối giản $\Leftrightarrow S(T)$ là mô tả được.

Chứng minh.

Phương pháp chứng minh như định lý 1

Định nghĩa 4

Giả sử $S(T) = [R, D_T, \sigma_T] \in S(TREE)$

Tập $L(S(T)) = [H(x) \mid H(x) \in R \wedge \exists A \subseteq D_T \wedge \sigma_T(H) = A]$ gọi là tập ngôn ngữ vào mô tả được trong $S(T)$.

ập $L_0(S(T)) = [H(x) \mid H(x) \in R \vee \exists d \in D_T \wedge \sigma_T(H) = [d]]$ gọi là tập nhân của
ập $L(S(T))$.

Chú ý $L_0(S(T)) \subseteq L(S(T))$

Định nghĩa 5

iả sử $H_1(x), H_2(x) \in L_0(S(T))$ với $S(T) = [R, D_T, \sigma_T]$.

$_1(x)$ là đồng dạng với $H_2(x)$ trên $L_0(S(T))$ (kí hiệu $H_1(x) \sqsupseteq H_2(x)$)

$\Leftrightarrow \exists d \in D_T \wedge \sigma_T(H_1) = \sigma_T(H_2) = [d]$.

ần

uan hệ $\sqsupseteq$ trên $L_0(S(T))$ là quan hệ tương đương và nó tạo ra một phân hoạch tương
ương. Mỗi lớp trong phân hoạch tương đương đó ta giữ lại một phần tử đại
iện và tập các phần tử đó ta kí hiệu là $L_0(S(T))/\sqsupseteq$. Rõ ràng là $L_0(S(T))/\sqsupseteq \subseteq L_0$
$S(T) \subset L(S(T))$.

Định lý 3

ới mỗi $(S(T))$ đều có tồn tại tập $L_0(S(T)/\sqsupseteq$ mô tả được trong $S(T)$

Chứng minh. Đối với T trong $S(T)$ có tồn tại cây chuẩn N (xem[7]). Dễ dàng
iểm tra lại $D_N = \cup \sigma_N(H)$, trong đó $L_0(S(N))/\sqsupseteq$ được xây dựng từ quá trình
$H \in L_0(S(N))/\sqsupseteq$

ây dựng dạng chuẩn N. Như vậy $L_0(S(N)) /\sqsupseteq$ là mô tả được trong $S(N)$ mà
$(T) \sim S(N)$ (Xem định lý 1) nên $L_0(S(N))/\sqsupseteq$ mô tả được trong $S(T)$.

Định nghĩa 6

iả sử $S(N) = [R, D_N, \sigma_N] \in S(TREE)$, N là cây chuẩn. Tập ngôn ngữ vào R
ược gọi là tập cốt yếu trong $S(N)$ nếu nó thỏa mãn các tính chất sau đây:

) R chỉ gồm các công thức cơ bản trong INEQU (tính cơ bản).

) $H \in R \leftrightarrow \exists d \in D_N \wedge \sigma_N(H) = [d]$ (tính mô tả được).

) Nếu bỏ đi khỏi R một công thức bất kỳ H thì tập $R \setminus [H]$ sẽ không mô tả được
rong $S(N)$ tính không thừa).

) Nếu thêm vào R một công thức $H \in FORM$ thì hoặc $H \in R$, hoặc $\exists H' \in R$ sao
ho $H \sqsupseteq H'$ hoặc $R \cup [H]$ sẽ không thỏa mãn một trong ba tính chất trên. (Tính
không thiếu).

Định lý 4

Nếu $N \in TREE$ là cây chuẩn thì tập $L_0(S(N))/\sqsupseteq$ là tập cốt yếu duy nhất
rong $S(N)$

Chứng minh

Kiểm tra lại 4 tính chất của tập cốt yếu. Tính duy nhất suy từ tính duy nhất của
lạng chuẩn N của T (xem trong [7]).

Định nghĩa 7

Hệ $S(T) = [R, D_T, \sigma_T]$ được gọi là tối giản nếu nó thỏa mãn

1) $S(T)$ là hệ D – tối giản

2) R là tập cốt yếu của $S(T)$.

Định lý 5

$\forall S(T) \in S(TREE) \exists$ hệ tối giản $S(T_0)$ sao cho $S(T) \sim S(T^0)$.

Chứng minh

Sử dụng các kết quả ở trên.

Chú ý: Dễ dàng chứng minh được nếu $S(T_1) \sim S(T_2)$ thì $S(T_1) \underset{R}{\sim} S(T_2)$. Điề ngược lại nói chung không đúng với mọi $R \subseteq FORM$. Nhưng nếu chọ $R_0 \in [L_0(S(N_1)) \rightleftharpoons, L_0(S(N_2))/\rightleftharpoons]$ thì ta có $S(T_1) \sim S(T_2). \leftrightarrow S(T_1) \sim S(T_2)$. Ở đâ N_i là cây chuẩn của T_i $(i = 1, 2)$.

TÀI LIỆU THAM KHẢO

1. G. Asser; Einführung in die mathematische logik. Teil 1, Leipzig, 1959.

2. W. Marek, Z. Pawlac, Bll. Acad. Polon., Sci., Sei. Math. Astronom. Phys. 22, 447. (1974)

3. W. Marek, Z Pawlak, Z Pawlak, Mathematical Foundations I, CC PA Reports, No 149, 1974

4. Salton G. – Automatic Information orgacisation and Retrieval New Yor 1968 (Mc Grow – Hill Book Company).

5. H. Thiele: Colloques Internationaux du C. N. R. S. No 296 (THEORIE DI L' INFORMATION 1977)

6. H. Thiele: Pr. IPI. Pan. 1980. No 411, 87 — 88.

7. Dp Duc Giao. Diss. A, Humboldt — Uni. Berlin, 1980.

8. Do Duc Giao: Pr. IPI. PAN. 1980. No 411. 33 — 35.

9. Đỗ Đức Giáo: «Thông báo khoa học 1983», Khoa Toán Cơ trường Đạ học Tổng hợp Hà nội, 1984, 64-70.

10. Đỗ Đức Giáo «Tạp chí khoa học» Toán Lý, số 2/1985, 22-26.

11. Đỗ Đức Giáo «Tóm tắt báo cáo Hội nghị toán học toàn quốc lần thứ thứ ị 22-25-1985), trang 97.

До Дык Зао, Ву Нгок Лоан

ОБ ПРОБЛЕМЕ МИНИМИЗАЦИИ СИСТЕМЫ РЕГУЛИРУЕМОЙ ИНФОРМАЦИИ

Трио $S(T) = [H, D_T, б_T]$ (T(–TREE) называется система регулируемо информации тогда и только тогда, когда R и D — непустве множества, а б преобразование из п в множество всех подмножеств множества D_T. Кажды элемент из R является фомулой и множество R называется множеством язы ков системы S(T). D_T есть множество documents присутствующиеся в дерев T и называется множеством documents системы S(T). $б_T$ — преобразовани регулирования системы S(T). Для кажзого элемента H —R, $б_T$ (H) – под множество documents b D_T. В этой работе рассмотривается проблема мини мизации системы регулируемой информации.

Do Duc Giao, Vu Ngoc Loan

MINIMIZATION PROPLEM FOR INFORMATION RETRIEVAL SYSTEM

A triple $S(T) = [R, D_T, \sigma_T]$ is called a information retrieval system if and only if R and D_T are non–empty sets and a mapping $\sigma_T : R \to 2^{D_T}$. The elements of R are the formulas and the Set R is said to be the language of S(T). D_T is the set of the documents in the tree T and the set D_T is interpreted as the document set of S(T). σ_T is calles the retrieval function of S(T) which give for each $H \in R$ the subset $\sigma_T(H) \subseteq D_T$ of documents retrieved by σ_T. In this paper we shall discuss the problem of minimization for the information retrieval systems consireded.

Nhận bài ngày 20-4-1986