Tạp chí Tin học và Điều khiển học, T. 12, S. 4 (1996) (125–136)

VỀ MỘT CÁCH TIẾP CẬN CƠ SỞ DỮ LIỆU THIẾU THÔNG TIN

TRƯƠNG ĐỨC HÙNG, LÊ TIẾN VƯƠNG

Abstract. This paper describes one approach to the problem of extended relational database.

Fuzzy data can be included in the system at two levels.

The first level considers the possibility of making fuzzy queries to the classic relational database.

The second level is related to the problem of adding fuzzy information to the system.

The evalution function τ and the threshold δ are introduced for handling and ordering the fuzzy information. On the extended domain of attribute we define different comparison operations are extended to the fuzzy relational database.

Cơ sở dữ liệu (CSDL) thiếu thông tin là một hướng được n hiều nhà nghiên cứu quan tâm phát triển và cài đặt. Một trong những cách tiếp cận có hiệu quả là sử dụng lý thuyết tập mờ và biến ngôn ngữ.

Trong khuôn khổ của bài này đưa ra một cách tiếp cận mới trên cơ sở đánh giá độ "khác biệt" giữa các giá trị ngôn ngữ làm nền tảng để đề xuất các phương pháp biểu diễn, lưu trữ và xử lý dữ liệu đồng thời có thể định nghĩa các phép tính đại số quan hệ mờ bằng cách mở rộng các phép tính đại số quan hệ thông thường nhờ lý thuyết tập mờ.

I - MỘT SỐ KHÁI NIỆM

Khái niệm tập mờ được L. A. Zadeh đưa ra từ 1965, trong đó mô tả các tính chất và các phép tính toán trên tập mờ. Phần này trình bày một số khái niệm cơ bản về tập mờ mà sẽ được áp dụng trong các phần tiếp theo (chi tiết có thể xem trong [20]).

Định nghĩa 1. Cho $U = \{u\}$ là vũ trụ các đối tượng xem xét. Tập mờ f trên U là tập các cặp có thứ tự $\{\mu_f(u), u\}$, với μ_f là hàm độ thuộc: $\mu_f : U \to [0, 1]$ gán cho mỗi phần tử $u \in U$ giá trị $\mu_f(u)$ phản ánh mức độ của u thuộc vào tập f. Để tiện trình bày cũng sẽ dùng ký hiệu sau để chỉ tập mờ:

$$f = \sum \mu_f(u)/u\,, \quad u \in U\,.$$

Giá đỡ (support) S_f của tập mờ f là tập các phần tử $u \in U$ có độ thuộc lớn hơn không: $S_f = \{u : \mu_f(u) > 0\}$.

Lõi (core) C_f của tập mờ f là tập các phần tử $u \in U$ có độ thuộc bằng một: $C_f = \{u : \mu_f(u) = 1\}$.

Tập mờ mức λ của f, ký hiệu f_λ, được định nghĩa bởi:

$$f_\lambda = \sum \mu_f(u)/u\,, \quad u \in f(\lambda)\,,$$

trong đó $f(\lambda)$ là tập mức λ của f và xác định như sau:

$$f(\lambda) = \{u : \mu_f(u) \geq \lambda\}\,, \text{ với } 0 \leq \lambda \leq 1\,.$$

Định nghĩa 2. Cho hai tập mờ: $f = \sum \mu_f(u)/u$ và $g = \sum \mu_g(u)/u$, $u \in U$.

a) f bằng g, ký hiệu $f = g$, nếu $\mu_f(u) = \mu_g(u)\ \forall u \in U$.

b) f chứa trong g, ký hiệu $f \subseteq g$, nếu: $\mu_f(u) \leq \mu_g(u)\ \forall u \in U$.

c) Bù của f: $\neg f = \sum(1 - \mu_f(u))/u\,,\ u \in U$.

d) Hợp của f và g: $f \cup g = \sum(\mu_f(u) \vee \mu_g(u))/u\,,\ u \in U$.

e) Giao của f và g: $f \cap g = \sum(\mu_f(u) \wedge \mu_g(u))/u\,,\ u \in U$.

f) Tổng hạn chế (bounded sum) của f và g:

$$f \oplus g = \sum(1 \wedge (\mu_f(u) + \mu_g(u))/u\,,\ u \in U.$$

g) Tích của f và g: $f * g = \sum \mu_f(u)\,.\,\mu_g(u)/u\,,\ u \in U$.

h) α là số thực, ký hiệu f^α là mũ bậc α của f: $f^\alpha = \sum \mu_f^\alpha(u)/u$.

i) Tích Đề các của $f_1, ..., f_n$ xác định tương ứng trên $U_1, ..., U_n$ là tập mờ trên $U_1 \times \cdots \times U_n$:

$$f_1 \times f_2 \times \cdots \times f_n \equiv \sum \mu_{f_1}(u_1) \wedge \cdots \wedge \mu_{f_n}(u_n)/(u_1, ..., u_n).$$

j) Nói f và g có liên quan với nhau nếu $S_f \cap S_g \neq \emptyset$, f và g có liên quan theo mức λ nếu $f(\lambda) \cap g(\lambda) \neq \emptyset$.

Ký hiệu $\wedge$ và $\vee$ là láy min và max.

Định nghĩa 3. Một biến ngôn ngữ được đặc trưng bởi bộ sáu

$(X, T(X), H, U, G, M)$, trong đó:

X - tên biến ngôn ngữ,

$T(X)$ hoặc đơn giản là T - tập các giá trị của biến ngôn ngữ X,

H - tập các gia tử,

U - vũ trụ các đối tượng xem xét hay còn gọi là tập cơ sở của biến X,

G - tập các quy tắc cú pháp sản sinh ra các giá trị ngôn ngữ,

M - tập các quy tắc ngữ nghĩa gán cho mỗi giá trị ngôn ngữ của biến X một ý nghĩa là tập mờ trên U.

Tập các giá trị ngôn ngữ $T(X)$, là miền trị của biến X, gồm số hữu hạn các giá trị nguyên thủy (primary term) và các giá trị phức hợp (composite term) được sản sinh nhờ tập các quy tắc cú pháp G.

Tập các quy tắc ngữ nghĩa M gán cho mỗi giá trị $t \in T$ một ý nghĩa $M(t)$ là tập mờ trên U. Ý nghĩa của các giá trị nguyên thủy được xác định tùy thuộc ngữ cảnh. Ý nghĩa của các giá trị phức xác định trên cơ sở tác động của các gia tử, các liên kết logic lên các giá trị nguyên thủy.

Khái niệm biến ngôn ngữ được áp dụng trong rất nhiều các lĩnh vực khác nhau. Trong CSDL mờ các thuộc tính của CSDL được coi là các biến ngôn ngữ cùng với miền trị của nó là tập các giá trị ngôn ngữ. Phần này đưa ra một số khái niệm và trình bày một số tính chất của biến ngôn ngữ cho phép xử lý dữ liệu ở dạng ngôn ngữ (trên tập $T(X)$) một cách thuận lợi, nhờ áp dụng các phép toán trên tập mờ.

Xét hai giá trị $t_1, t_2 \in T$.

1. Nói t_1 và t_2 là bằng nhau, ký hiệu $t_1 = t_2$, nếu $M(t_1) = M(t_2)$.

Tương tự có $t_1 \overset{\lambda}{=} t_2$ nếu $M(t_1) \overset{\lambda}{=} M(t_2)$.

2. Nói t_1 và t_2 có quan hệ với nhau nếu $S_{M(t_1)} \cap S_{M(t_2)} \neq \emptyset$, t_1 và t_2 có quan hệ với nhau theo mức λ nếu $M(t_1)(\lambda) \cap M(t_2)(\lambda) \neq \emptyset$, trong đó $M(t_1)(\lambda)$ và $M(t_2)(\lambda)$ là tập mức λ tương ứng của $M(t_1)$ và $M(t_2)$.

3. t_1 chứa trong t_2, ký hiệu $t_1 \subseteq t_2$, nếu $M(t_1) \subseteq M(t_2)$; t_1 chức trong t_2 theo mức λ, $t_1 \overset{\lambda}{\subseteq} t_2$, nếu $M(t_1) \overset{\lambda}{\subseteq} M(t_2)$.

Quan hệ $\subseteq$ tạo thành một quan hệ thứ tự bộ phận:

- $t_1 \subseteq t_2$ và $t_2 \subseteq t_3 \Rightarrow t_1 \subseteq t_3$.
- $t_1 \subseteq t_2$ và $t_2 \subseteq t_1 \Rightarrow t_1 = t_2$.

4. Độ khác biệt giữa t_1 và t_2, ký hiệu $\rho(t_1, t_2)$, được xác định như một ánh xạ $\rho : T \to [0, 1]$ như sau:

- $\rho(t, t) = 0$, $\forall t \in T$.
- $\rho(t_1, t_2) = \rho(t_2, t_1) \leq 1$, $\forall t_1, t_2 \in T$.
- $\rho(t_1, t_2) = 1$ nếu t_1 và t_2 không có quan hệ với nhau. Trong các phần tiếp theo sẽ sử dụng ρ định nghĩa như sau:

$$\rho(t_1, t_2) = \begin{cases} \dfrac{\sqrt{\sum_{u \in S_{t_1} \cup S_{t_2}} (\mu_{t_1}(u) - \mu_{t_2}(u))^2}}{|S_{t_1} \cup S_{t_2}|} & \text{nếu } S_{t_1} \cap S_{t_2} \neq \emptyset \\ 1 & \text{nếu } S_{t_1} \cap S_{t_2} = \emptyset \end{cases}$$

trong đó $|\cdot|$ ký hiệu lực lượng của tập.

5. Nói t_1 và t_2 là δ-tương đương, ký hiệu $t_1 \overset{\delta}{\approx} t_2$ (hoặc ngắn gọn là $t_1 \approx t_2$), nếu $\rho(t_1, t_2) \leq \delta$, với $0 \leq \delta \leq 1$.

Trên đây là các khái niệm quan trọng (nhất là khái niệm 4 và 5) mà sẽ được sử dụng ở các phần sau, đặc biệt để xây dựng các phương pháp đánh giá và xử lý dữ liệu.

II - CƠ SỞ DỮ LIỆU MỞ RỘNG

Trong những năm gần đây việc nghiên cứu mở rộng mô hình CSDL quan hệ truyền thống để có thể bao hàm được các thông tin không đầy đủ ngày càng được nhiều người quan tâm. Dữ liệu với thông tin không đầy đủ trong CSDL được xem xét trên hai khía cạnh: biểu diễn, lưu trữ và xử lý.

Trong [5, 12, 13] đã xem xét việc mở rộng câu hỏi tìm kiếm thông tin trong CSDL quan hệ để có thể xử lý được một số dạng dữ liệu mờ. Việc đưa dữ liệu mờ vào CSDL đã dẫn đến việc mở rộng mô hình CSDL truyền thống thành mô hình CSDL mờ. Trong [7] đã đưa ra mô hình khá tổng quát của CSDL. J. M. Medina và O.Pons [9], đã đề xuất mô hình mở rộng của CSDL trong đó đề cập đến việc định nghĩa cấu trúc dữ liệu mờ với công cụ xử lý và một vài khía cạnh về cài đặt. Tuy nhiên việc mở rộng CSDL gặp không ít khó khăn trong xử lý dữ liệu mờ.

Tiếp tục cách tiếp cận trong [7] bài báo này mở rộng việc đánh giá dữ liệu: đánh giá tương đương, lớn hơn, nhỏ hơn giữa các giá trị mờ và như vậy sẽ mở rộng một cách đầu đủ các phép tính quan hệ. Các phương pháp đánh giá ở đây tương đối tự nhiên và dễ dàng trong cài đặt.

1. Định nghĩa và biểu diễn dữ liệu

Ký hiệu R (hoặc $R(W)$) là quan hệ trên tập thuộc tính $W = \{A_1, ..., A_n\}$. Mỗi thuộc tính $A \in W$ được xem như là một biến ngôn ngữ có miền trị cơ sở $U(A)$ bao gồm những giá trị được xác định rõ của A. Miền trị $U(A)$ có thể được mở rộng bằng các giá trị ngôn ngữ các tập mờ để có thể được các mô tả gần đúng. Như vậy miền trị của A, ký hiệu $D(A)$, sẽ là hợp của miền trị cơ sở và miền trị mở rộng

$$D(A) = U(A) \cup T(A) \cup F(A),$$

trong đó:

$T(A)$ - tập giá trị ngôn ngữ của biến A,

$F(A)$ - tập mờ trên miền trị cơ sở $U(A)$.

Quan hệ R với miền trị xác định như trên gọi là quan hệ được mở rộng (hoặc đơn giản là quan hệ mở rộng): $R \subseteq D(A_1) \times \cdots \times D(A_n)$.

Định nghĩa 4. Cơ sở dữ liệu được mở rộng (hoặc đơn giản là CSDL mở rộng) DB là tập hữu hạn các quan hệ mở rộng: $DB = \{R_1, ..., R_n\}$, mỗi R_i xác định

trên tập hữu hạn các thuộc tính $W_i = \{A_{i1}, ..., A_{ik}\}$:

$$R_i \subseteq D(A_{i1}) \times \cdots \times D(A_{ik})$$

Xét quan hệ $R_i \subseteq D(A_{i1}) \times \cdots \times D(A_{ik})$, xây dựng các ánh xạ:

$$M_i : D(A_i) \to M(A_i),\ i = 1, ..., n$$

Gán cho mỗi $t \in D(A_i)$ một ý nghĩa $m(t) \in M(A_i)$ như sau:

1. Nếu $t \in U(A_i)$ thì $m(t) = 1/t$.
2. Nếu $t \in F(A_i)$ thì $m(t) = t = \sum \mu_t(u)/u$.
3. Nếu t $t \in T(A_i)$ thì $m(t)$ là tập mờ xác định ngữ nghĩa của giá trị ngôn ngữ t (Theo quy tắc ngữ nghĩa của biểu ngôn ngữ).

Ánh xạ M_i cho phép biểu diễn dữ liệu của thuộc tính A_i dưới dạng chung và đưa việc xử lý dữ liệu trên tập D(Ai) về dạng xử lý tương đương trên tập $M(A_i)$. Ký hiệu $M = (M_1, ..., M_n)$ là ánh xạ bậc n từ tập $D(A_1) \times \cdots \times D(A_n)$ sang tập $M(A_1) \times \cdots \times M(A_n)$, khí đó đối với mỗi $R \subseteq D(A_1) \times \cdots \times D(A_n)$ có $R^* = M(R) \subseteq M(A_1) \times \cdots \times M(A_n)$. R được gọi là biểu diễn ngoài còn R^* là biểu diễn trong của quan hệ; r là biểu diễn ngoài, r^* là biểu diễn trong của một bộ; $D(A)$ là biểu diễn ngoài, $M(A)$ là biểu diễn trong của miền trị thuộc tính A.

Đối với mỗi quan hệ R dữ liệu có ba dạng: hằng $u \in U(A_i)$; các tập mờ $f \in F(A_i)$ và các giá trị ngôn ngữ $t \in T(A_i)$. Ngược lại đối với quan hệ R^* dữ liệu chỉ có một dạng duy nhất là các tập mờ.

Để đơn giản cho trình bày sẽ dùng chung ký hiệu $D(A)$ để chỉ miền trị của thuộc tính A ở cả biểu diễn trong và biểu diễn ngoài.

Trong biểu diễn ngoài của quan hệ có thể có các giá trị ngôn ngữ trùng nhau, nhưng ở biểu diễn trong chúng lại là các giá trị phân biệt.

Biểu diễn trong của quan hệ thuận lợi cho việc xử lý dữ liệu, tạo điều kiện cho việc xây dựng phương pháp xử lý dữ liệu chung cho mọi dạng dữ liệu trong CSDL (dữ liệu rõ và mờ). Tuy nhiên việc lưu trữ biểu diễn trong R^* rất cồng kềnh và tốn kém. Trong các ứng dụng cụ thể chỉ cần lưu trữ một phần rất nhỏ của R^*. Để thực hiện điều đó tiến hành như sau:

Ký hiệu $T^0(A) = (t_1^0, ..., t_k^0)$ là tập các giá trị nguyên thủy của $T(A)$. Rõ ràng, $T^0(A)$ là hữu hạn và có lực lượng rất nhỏ so với $T(A)$.

Ký hiệu $H(A)$ là tập các gia tử của A và $M^0(A)$ là lõi của biểu diễn trong $M(A)$: $M^0(A) = \{m(t) : t \in U(A) \cup T^0(A)\}$.

Rõ ràng là bất kỳ giá trị nào của $M(A)$ cũng có thể thu đươc thông qua các giá trị của $M^0(A)$. Như vậy thay vì phải lưu giữ toàn bộ biểu diễn trong $M(A)$ chỉ cần lưu giữ lõi của nó là $M^0(A)$.

Việc lưu giữ tập các gia tử $H(A)$ phụ thuộc vào định nghĩa $h \in H(A)$ trong các ứng dụng cụ thể. Nhìn chung mỗi $h \in H(A)$ gắn với mỗi thủ tục sinh ra $h.t$ khi tác động lên t.

2. Xử lý dữ liệu

Để định nghĩa các phép toán quan hệ trong CSDL mờ cần xem xét các phép so sánh, đánh giá các giá trị trong miền D của thuộc tính.

Trong [4,7,13] hai giá trị $f, g \in D$ được coi là bằng nhau nếu ý nghĩa của chúng bằng nhau, trên cơ sở đó mở rộng phép toán chọn và kết nối bằng. Trong [11,16] đã dùng quan hệ gần gũi để mở rộng khái niệm bằng nhau giữa hai giá trị mờ. Tiếp tục kết quả của các tác giả trên, trong bài báo này đề xuất một tiêu chuẩn đánh giá làm cơ sở xác định các quan hệ "tương đương" ($\approx$), "đứng trước" ($\langle$), "đứng sau" ($\rangle$), "khác nhau" ($\not\approx$) trên miền trị D theo nghĩa sẽ trình bày dưới đây.

Ký hiệu $\tau(f\,\theta\,g)$ là hàm đánh giá f, g thỏa $f\,\theta\,g$, trong đó $f, g \in D$ với $\theta \in \{\approx, \langle, \langle\approx, \rangle, \rangle\approx, \not\approx\}$, $\tau(f\,\theta\,g) \in [0, 1]$.

Định nghĩa 5. Cho thuộc tính A với miền trị D, $f, g \in D$.

Hai giá trị f, và g gọi là tương đương ($f \approx g$) nếu $\tau(f \approx g) > 0$, trong đó $\tau(f \approx g)$ được xác định là: $\tau(f \approx g) = 1 - \rho(f, g)$, với $\rho(f, g)$ là độ khác biệt giữa f và g. Nếu $\tau(f \approx g) = 1$ ($\rho(f, g) = 0$) thì $f = g$ theo nghĩa bằng nhau giữa các tập mờ.

Từ Định nghĩa 5 trực tiếp suy ra:

Hệ quả 1.

$\tau(f \approx f) = 1\ \forall f \in D$; $\tau(f \approx g) = \tau(g \approx f)\ \forall f, g \in D$.

$\tau(f \approx g) = 0$ *nếu* $S_f \cap S_g = \emptyset$.

Để xây dựng quan hệ "trước, sau" giữa các giá trị trong D của thuộc tính mà có miền trị cơ sở U là tập thứ tự; xét $f, g \in D$:

$$f = \sum \mu_f(u)/u\,, \quad g = \sum \mu_g(u)/u\,.$$

Ký hiệu $M_f = \{u : \mu_f(u) = \max_{u' \in U}(\mu_f(u'))\}$,

$\mathrm{Max}\,(f) = \max_{u \in M_f}(u)$, $\mathrm{Max}\,(g) = \max_{u \in M_g}(u)$.

Trong [4,6] đưa ra phương pháp sắp xếp dựa trên miền giá đỡ của các giá trị, với đánh giá G phát biểu rằng giá trị g đứng trước giá trị f (ký hiệu $g \leq f$) khi và chỉ khi $G(\geq g/f) = \mathrm{TRUE}$ và $G(\leq f/g) = \mathrm{TRUE}$, nghĩa là $S_f \subseteq S_{\geq g}$ và $S_g \subseteq S_{\leq f}$. Theo phương pháp này f và g không phải luôn luôn xếp thứ tự "trước, sau" được với nhau vì cùng một lúc có $f \leq_G g$ và $g \leq_G f$. ($f \leq_G g$ ký hiệu f đứng trước g theo đánh giá G). $s>g$

L. T. Koczy [6] đề xuất phương pháp sắp xếp các tập mờ CNF (tập chuẩn lồi) bằng cách so sánh các min và max của miền giá đỡ: tập g đứng trước f ($g\langle f$) khi và chỉ khi $\min(g) < \min(f)$ và $\max(g) < \max(f)$. Phương pháp này cũng có nhiều hạn chế.

Ở đây xây dựng hàm đánh giá dựa trên hai yếu tố: hàm độ thuộc và tập giá đỡ.

- Mỗi giá trị $f, g \in D$ được biểu diễn dưới dạng đồ thị của hàm độ thuộc. Thứ tự của giá trị cơ sở u là một yếu tố quan trọng giúp cho việc sắp xếp "trước, sau" giữa f và g.

- Yếu tố thứ hai là xét trên mức độ mang thông tin của hai giá trị. Nếu hai giá trị có miền giao nhau thì giá trị mang ít thông tin hơn là giá trị đứng trước, giá trị mang nhiều thông tin hơn là giá trị đứng sau.

Trên cơ sở lập luận trên, thiết lập hai hàm cho các giá trị f, g như sau:

$$\tau^1(f, g) = \begin{cases} 1 & \text{nếu } \mathrm{Max}(f) > \mathrm{Max}(g) \\ 0 & \text{nếu } \mathrm{Max}(f) = \mathrm{Max}(g) \\ -1 & \text{nếu } \mathrm{Max}(f) < \mathrm{Max}(g) \end{cases}$$

$$\tau^2(f, g) = \frac{\sum_{u \in S_f \cup S_g} \mu_f(u) - \mu_g(u)}{|S_f \cup S_g|}$$

Định nghĩa 6. Hàm đánh giá thứ tự giữa hai giá trị f và g được định nghĩa qua:

$$\tau(f, g) = \max(\min(\tau^1(f, g) + \tau^2(f, g), 1), 0).$$

- Nếu $\tau(f, g) > 0$ nói rằng f đứng sau g (f lớn hơn g), ký hiệu $f \rangle g$, với giá trị hàm đánh giá $\tau(f \rangle g) \equiv \tau(f, g)$.

- Nếu $f \rangle g$ thì nói rằng $g \langle f$ với giá trị hàm đánh giá $\tau(g \langle f) \equiv \tau(f \rangle g)$.

Như sẽ trình bày dưới đây hàm đánh giá τ là tiêu chuẩn khá tốt để sắp xếp tập giá trị D của thuộc tính với miền cơ sở U có thứ tự. Trong trường hợp $f, g \in D$ là giá trị rõ thì $f \rangle g$ cho kết quả như phép sánh thông thường $f > g$.

Định lý 1. *Cho $f, g, h \in D$. Nếu $f \rangle g$ và $g \rangle h \Rightarrow f \rangle h$.*

Chứng minh. Dựa vào định nghĩa 6, so sánh hàm độ thuộc của f và g trên hợp của miền giá đỡ tương ứng sẽ cho kết quả.

Định nghĩa 7.

$f \not\approx g$ nếu f không tương đương g.

$f \rangle \approx g$ nếu $f \rangle g$ hoặc $f \approx g$ với $\tau(f \rangle \approx g) = \max(\tau(f \rangle g), \tau(f \approx g))$.

$f \langle \approx g$ nếu $f \langle g$ hoặc $f \approx g$ với $\tau(f \langle \approx g) = \max(\tau(f \langle g), \tau(f \approx g))$.

Hàm đánh giá của phủ định KHÔNG (ký hiệu $\rceil$) định nghĩa qua:

$$\tau(\rceil(f \rangle g)) = \tau(f \langle \approx g), \quad \tau(\rceil(f \langle g)) = \tau(f \rangle \approx g).$$

Ký hiệu p, q là các tân từ mờ dạng $f \theta g$, với $f, g \in D$,

$\theta \in \{\approx, \langle, \langle\approx, \rangle, \rangle \approx\}$.

Hàm đánh giá của p Và q, p Hoặc q được xác định qua

$\tau(p \text{ Và } q) = \tau(p) \wedge \tau(q)$, $\tau(p \text{ Hoặc } q) = \tau(p) \vee \tau(q)$,

trong đó $\wedge$ và $\vee$ ký hiệu Min và Max.

Nếu R-quan hệ mở rộng trên W; $r_1, r_2 \in R$; $A \in W$ thì:

1. Hai bộ r_1 và r_2 gọi là tương đương trên thuộc tính A, ký hiệu $r_1 \approx_A r_2$ nếu $r_1[A] \approx r_2[A]$.

2. Hai bộ r_1, r_2 tương đương trên $X \subseteq W$, ký hiệu $r_1 \approx_X r_2$, nếu $\forall A \in X$, $r_1 \approx_A r_2$.

3. Quan hệ R_1 và R_2 gọi là tương trên $X \subseteq W$, ký hiệu $R_1 \approx_X R_2$, nếu $\forall r_1 \in R_1$, $\exists r_2 \in R_2$ sao cho $r_1 \approx_X r_2$ và $\forall r_2 \in R_2$, $\exists r_1 \in R_1$ sao cho $r_2 \approx_X r_1$.

4. Giá trị chân lý của $r_1[X]\,\theta\, r_2[X]$ với $X \subseteq W$ và $\theta \in \{\approx, \langle, \langle\approx, \rangle, \rangle \approx\}$ được tính như sau:

$$\tau(r_1[X]\,\theta\, r_2[X]) = \bigwedge_{A \in X} \tau(r_1[A]\,\theta\, r_2[A])$$

4. Các phép toán quan hệ

Trong [7, 4] các tác giả đã nghiên cứu, mở rộng các phép toán quan hệ của CSDL truyền thống cho CSDL mở rộng, trong đó đánh giá thỏa của các giá trị khi thực hiện phép chọn và kết nối được mở rộng từ khái niệm bằng thông thường thành khái niệm bằng giữa các tập mờ.

Tiếp tục phát triển tiếp cận trong [7], trên cơ sở các quy tắc xử lý dữ liệu trình bày ở trên, phần này mở rộng theo hai mô hình một cách đầy đủ các phép toán quan hệ cho CSDL mở rộng và CSDL mờ.

Tập các phép sánh $\{\approx, \langle, \langle\approx, \rangle, \rangle \approx\}$ là mở rộng của tập các phép sánh thông thường $\{=, <, \leq, >, \geq\}$ trên U.

Giả thiết rằng ngưỡng $\delta \in [0, 1]$ được chọn các đánh giá thỏa: f và g thỏa $f\,\theta\, g$ nếu $\tau(f\,\theta\, g) \geq \delta$.

Phép δ-chọn

Cho quan hệ R xác định trên W và $A \in W$, $t \in D(A)$. Phép δ-chọn các bộ trong R thỏa mãn điều kiện cho $A\,\theta\, t$ định nghĩa như sau:

$$SL_{A\,\theta\, t}(R) = \{r : r \in R,\ \tau(r[A]\,\theta\, t) > \delta\}$$

trong đó $\theta \in \{\approx, \langle, \langle\approx, \rangle, \rangle \approx\}$; δ-ngưỡng chọn.

Phép δ-kết nối

Cho hai bộ $r1 = (f_1, ..., f_n)$ và $r_2 = (g_1, ..., g_n)$ khi đó ký hiệu $(r_1, r_2) = (f_1, ..., f_n, g_1, ..., g_n)$.

Giả sử R_1 và R_2 là hai quan hệ trên W_1 và W_2, $A \in W_1$, $B \in W_2$ là hai thuộc tính cùng miền trị. Phép σ-kết nối của R_1 và R_2 trên A và B được định nghĩa

như sau:

$R_1[A\,\theta\,B]R_2 = \{(r_1, r_2) : r_1 \in R_1,\ r_2 \in R_2,\ \tau(r_1[A]\,\theta\, r_2[B] \geq \delta\}$

Trường hợp θ là $\approx$, phép δ-kết nối gọi là kết nối tương đương. Khi kết nối tương đương tại thuộc tính cùng tên của hai quan hệ thì được phép kết nối tự nhiên:

$R_1 * R_2 = \{r : \exists r_1 \in R_1,\ r_2 \in R_2$, sao cho $r[X] = r_1[X]$ và $r[Z] = r_2[Z]$ và $r[y] = r_1[y] \approx r_2[y]\}$

trong đó R_1 xác định trên XY; R_2 - trên YZ, với X, Y, Z ký hiệu các tập thuộc tính rời nhau. Ở đây $r_1[y] \approx r_2[y]$ nếu $\tau(r_1[y] \approx r_2[y]) \geq \delta$.

Phép δ-chiếu

Giả sử R là quan hệ mở rộng trên W và $X = \{A_1, ..., A_k\} \subseteq W$.

Phép δ-chiếu của R trên X định nghĩa như sau:

$PR_X[R] = \{$gồm các bộ $r[X]$ sao cho: $\exists r' \in R$ mà $r'[X] \approx r[X]$, $\exists r'' \in R$ mà $r''[X] \approx r[X]$ và $r''[X] \not\approx r'[X]\}$

Như vậy $PR_X(R)$ là quan hệ nhận được từ R bằng cách xóa bỏ tất cả các cột không thuộc X và thay thế các bộ tương đương nhau trong $PR_X(R)$ bằng một đại diện của chúng.

Phương pháp xử lý dữ liệu cùng các phép toán quan hệ mở rộng trình bày trên đây so với các nghiên cứu trước đó (như [7,4,6,8,9,12,13]) có các ưu điểm sau:

- Biểu thức chọn trong các phép toán quan hệ được mở rộng với $\theta \in \{\approx, \langle, \langle\approx, \rangle, \rangle \approx\}$ cho phép biểu diễn và xử lý các câu hỏi tìm kiếm gần gũi với ngôn ngữ tự nhiên và với các yêu cầu ở các mức độ phức tạp khác nhau cho cả dữ liệu rõ và dữ liệu mờ.

- Kết quả của các phép toán quan hệ mở rộng không phải chỉ gồm các bộ hoàn toàn không liên quan đến nhau mà gồm các tập các bộ "gần gũi" nhau theo hàm đánh giá τ và ngưỡng chọn δ. Việc chọn bộ "thỏa đáng nhất" từ tập các bộ "gần gũi" nhau do người sử dụng thực hiện dựa trên các thông tin hỗ trợ khác.

Một kiểu mở rộng tiếp tục cho quan hệ mờ được xác định như sau:

Định nghĩa 8. Cho $W = \{\mu, A_1, ..., A_n\}$ là tập các thuộc tính và $D(\mu)$, $D(A_1)$, ..., $D(A_n)$ là các miền trị tương ứng, trong đó $D(\mu) = [0, 1]$.

Cơ sở dữ liệu mờ DF trên W là tập các quan hệ mờ FR:

$$DF = (FR_1, ..., FR_m)$$

FR là một quan hệ mở rộng trên W: $FR \subseteq D(\mu) \times D(A_1) \times \cdots \times D(A_n)$.

Thuộc tính μ là thuộc tính đặc biệt gọi là độ thuộc (thuộc tính độ thuộc) với miền trị là đoạn $[0, 1]$.

Ký hiệu r là bộ của quan hệ R trên $D(A_1) \times \cdots \times D(A_n)$, bộ của quan hệ

FR được ký hiệu là $(\mu_{FR}(r), r)$, trong đó $\mu_{FR}(r) \in [0, 1]$ là giá trị của bộ r tại μ (thay vì ký hiệu $r[\mu]$ dùng ký hiệu $\mu_{FR}(r)$). Vậy quan hệ mờ FR được biểu diễn:

$FR = \{(\mu_{FR}(r), r) : \mu_{FR}(r) \in [0, 1],\ r \in D(A_1) \times \cdots \times D(A_n)\}$

Đôi khi cũng sẽ dùng ký hiệu ngắn gọn là:

$FR = \{(\mu_{FR}(r), r) : r \in R\}$ hoặc $FR = \{(\mu_{FR}(r), r)\}$

Giá trị $\mu_{FR}(r)$ xác định mức độ bộ r xuất hiện trong quan hệ FR. Trong ứng dụng cụ thể bọ $r \in FR$ khi và chỉ khi $\mu_{FR}(r) \geq \delta$ với δ là ngưỡng được chọn. Mỗi quan hệ mở rộng R cũng là một quan hệ mờ nếu coi mỗi bộ $r \in R$ có độ thuộc bằng 1: $R = \{(1, r)\}$.

Các phép toán cơ bản của quan hệ mờ được định nghĩa như sau:

Các phép toán tập hợp trên các quan hệ mờ

1. Giả sử FR_1 và FR_2 là hai quan hệ mờ xác định trên cùng tập thuộc tính W.

- Hợp của FR_1 và FR_2 là:

$$FR_1 \cup FR_2 = \{(\mu_{FR_1}(r) \vee \mu_{FR_2}(r), r) : (\mu_{FR_1}(r), r) \in FR_1 \text{ hoặc } (\mu_{FR_2}(r), r) \in FR_2\}$$

- Giao của FR_1 và FR_2 là:

$$FR_1 \cap FR_2 = \{(\mu_{FR_1}(r) \wedge \mu_{FR_2}(r), r) : (\mu_{FR_1}(r), r) \in FR_1 \text{ và } (\mu_{FR_2}(r), r) \in FR_2\}$$

2. Nếu FR_1 xác định trên W_1 và FR_2 xác định trên W_2 thì tích Đề - các của FR_1 và FR_2 là:

$$FR_1 \times FR_2 = \{(\mu_{FR_1}(r_1) \wedge \mu_{FR_2}(r_2), (r_1, r_2) : (\mu_{FR_1}(r_1), r_1) \in FR_1, (\mu_{FR_2}(r_2), r_2) \in FR_2\}$$

Phép chọn mờ

Cho FR là quan hệ mờ trên W; $A \in W$, $t \in D(A)$. Phép chọn mờ FSL các bộ trong FR thỏa điều kiện $A\,\theta\,t$ định nghĩa như sau:

$FSL_{A\,\theta\,t}(FR) = \{(\mu_{FR}(r) \wedge \tau(r[A]\,\theta\,t), r) : (\mu_{FR}(r), r) \in FR\}$.

Bộ $(\mu_{FR}(r) \wedge \tau(r[A]\rho t), r) \in FSL$ khi và chỉ khi $\mu_{FR}(r) \wedge \tau(r[A]\,\theta\,t) \geq \delta$.

Phép chiếu mờ

Cho quan hệ mờ $FR = \{(\mu_{FR}(r), r)\}$ và $X \subseteq W$.

Với mỗi bộ $r \in R$ ký hiệu r_x là tập các bộ $r' \in R$ mà có giá trị tại X tương đương với giá trị r tại X: $r_x = \{r' \in R : r'[X] \approx r[X]\}$.

Phép chiếu của FR trên X định nghĩa như sau:

$$FPR_X(FR) = FPR[X] = \{(\bigvee_{r' \in r_x} \mu_{FR}(r'), r[X])\}$$

Vậy trong quan hệ kết quả của phép chiếu mờ các bộ tương đương được đồng nhất với bộ có độ thuộc lớn nhất.

Phép kết nối mờ

Cho $FR_1 = \{(\mu_{FR_1}(r_1), r_1)\}$ và $FR_2 = \{(\mu_{FR_2}(r_2), r_2)\}$ là hai quan hệ mờ trên W_1 và W_2; $A \in W_1$, $B \in W_2$.

Phép kết nối mờ của FR_1 và FR_2 trên A và B định nghĩa như sau:

$FR_1[A\,\theta\,B]FR_2 \equiv FR = \{(\mu_{FR}(r_1, r_2), (r_1, r_2)) : (\mu_{FR_1}(r_1), r_1) \in FR_1$ và $(\mu_{FR_2}(r_2), r_2) \in FR_2$ và $r_1[A]\,\theta\,r_2[B]\}$

trong đó: $\mu_{FR}(r_1, r_2) = \min(\mu_{FR_1}(r_1), \mu_{FR_2}(r_2), \tau(r_1[A]\,\theta\,r_2[B])$

Trường hợp θ là $\approx$ phép kết nối gọi là kết nối mờ tương đương. Nếu nối tại thuộc tính cùng tên của hai quan hệ thì được kết nối mờ tự nhiên.

$FR_1 * FR_2 \equiv FR = \{(\mu_{FR}(r), r) : \exists(\mu_{FR_1}(r_1), r_1) \in FR_1$ và $(\mu_{FR_2}(r_2), r_2) \in FR_2$ sao cho $r[X] = r_1[X]$ và $r[Z] = r_2[Z]$ và $r[Y] = r_1[Y] \approx r_2[Y]\}$

Trong đó $\mu_{FR} = \min(\mu_{FR_1}(r_1), \mu_{FR_2}(r_2), \tau(r_1[Y] \approx r_2[y]))$; FR_1 xác định trên XY; FR_2 trên YZ với X, Y, Z là tập thuộc tính rời nhau.

KẾT LUẬN

Mô hình CSDL mở rộng được phát triển trên cơ sở mô hình CSDL truyền thống bằng các công cụ của lý thuyết tập mờ và biến ngôn ngữ. Cấu trúc của dữ liệu được mở rộng để tiếp nhận cả dữ liệu xác định và dữ liệu mờ, nhờ đó đã đề xuất công cụ xử lý dữ liệu dựa trên hàm đánh giá, mà một mặt khá đơn giản, đáp ứng được cách hiểu trực quan đối với dữ liệu rõ, dễ dàng cài đặt, mặt khác cho phép đánh giá, phân hoạch miền trị đối với thuộc tính mà miền trị cơ sở là hữu hạn. Các phép toán quan hệ và quan hệ mờ là mở rộng đầy đủ của các phép toán quan hệ truyền thống, cho CSDL mờ.

Mô hình ở dạng đơn giản đã được cài đặt trong khuôn khổ đề tài cấp Nhà nước mã số KC-01-03 và đã được nghiệm thu. Thiết kế hoàn chỉnh của mô hình được cài đặt trên MS-ACCESS. Chương trình được xây dựng trên ngôn ngữ Visual Basic, đòi hỏi hệ thống có cấu hình tối thiểu là máy 486 với 8MB RAM trở lên và môi trường Windows. Mô hình được thử nghiệm áp dụng cho hệ thống thông tin truy tìm tội phạm.

Các tác giả xin chân thành cảm ơn Ban chủ nhiệm Đề tài KC-01-03 đã tạo điều kiện thuận lợi cho việc thử nghiệm và cài đặt mô hình này, đồng thời các tác giả cũng bày tỏ sự biết ơn sâu sắc đến Ban biên tập Tạp chí Tin học và Điều khiển học đã có những ý kiến quý báu và tạo điều kiện để công bố kết quả này.

TÀI LIỆU THAM KHẢO

1. B. P. Buckles and Petry, *A Fuzzy representation of data for relational database*, Fuzzy sets and systems, **7** (1982), 213 - 226.

2. B. P. Buckles and F. E. Petry, *Extending the fuzzy database with fuzzy numbers*, Inform Sci., **34** (1984), 145 - 155.
3. E. E. Codd, *Extending database relational model to capture more meaning*, ACM Trans. Database system, **41** (1979).
4. Đinh Thị Ngọc Thanh, *CSDL mở rộng với thông tin không đầy đủ*, Luận án phó tiến sỹ, Hà Nội, 1991.
5. Hệ quản trị CSDL Famebase ver. 3.0, Tài liệu đề tài cấp Nhà nước mã số 90-68-051.
6. L. T. Koczy, K. Hirota, *Ordering distance and closeness of fuzzy sets*, Fuzzy sets and systems, **59** (1993), 281 - 293.
7. Le Tien Vuong, Ho Thuan, *A relational database extended by application of fuzzy set theory and liguistic variables*, Computers and Artificial Intelligence, **8** (2) (1989).
8. Le Tien Vuong, Ho Thuan, *Retrival from fuzzy database by fuzzy relational algebra*, MTA SZTAKI Kozlemenyek, **37** (1987).
9. J. M. Medina, O. Pons, and M. A. Vila, *Gefred - A generalized model of fuzzy relational database*, Inform. Sci., *76* (1994), 87 - 109.
10. P. C. Saxena, B. K. Tyaga, *Fuzzy functional dependencies and undependencies in extended fuzzy relational database models*, Fuzzy sets and systems, **69** (1995), 65 - 89.
11. S. Shenoi, A. Melton, *Proximity relations in the fuzzy relational database model*, Fuzzy sets and system, **31** (1989), 285 - 296.
12. S. Shenoi and A. Melton, *A extended version of the fuzzy relational database model*, Inform. Sci., **52** (1990), 35 - 52.
13. Thiết kế và cài đặt hệ lập luận xấp xỉ trong CSDL thiếu thông tin, Tài liệu cấp nhà nước, mã số KC-01-03, 1993.
14. Trương Đức Hùng, Lê Tiến Vương, *Thiết kế và cài đặt cơ chế lập luận xấp xỉ trong hệ quản trị CSDL thiếu thông tin*, Tài liệu Hội nghị vô tuyến Điện tử, Hà Nội, 1992.
15. Trương Đức Hùng, Lê Tiến Vương, ıBiểu diễn dữ liệu và câu hỏi tìm kiếm với cơ chế lập luận xấp xỉ, Tài liệu tuần lễ tin học lần IV, 8/1994.
16. Trương Đức Hùng, Lê Tiến Vương, *Xấp xỉ ngôn ngữ trong CSDL thiếu thông tin*, Tài liệu Tuần lễ Tin học lần IV, 8/1994.
17. Trương Đức Hùng, *Một phương pháp xử lý dữ liệu trong CSDL mờ*, Tài liệu Hội nghị khoa học Trường đại học Bách khoa Hà Nội nhân dịp 40 năm thành lập trường, 10/1996.
18. M. Umano, M. Mizumoto, *FSTDS-systems: a fyzzy set manipulation system*, Inform. Sci., **14** (1978) 115 - 159.
19. L. A. Zadeh, *The concept of linguistic variable and its applications to approximate reasoning*, Inf. Sci., **8** (1975), 199 - 248 and 301 - 357.
20. L. A. Zadeh, *Fuzzy sets*, Inf. and Contr., **8** (1965), 338 - 353.

Cục khoa học Viễn thông - Tin học, Bộ Nội vụ.
Viện Điều tra Quy hoạch đất đai.

Nhận bài ngày 16-4-1996