VỀ ÁP DỤNG LÝ THUYẾT TẬP MỜ TRONG MÔ HÌNH QUAN HỆ

LÊ TIẾN VƯƠNG

Sự phát triển nhanh chóng kỹ thuật máy tính điện tử và việc cài đặt các hệ thống chương trình đặc biệt đã tạo khả năng lớn cho việc sử dụng và quản lý khối lượng thông tin lớn dưới dạng ngân hàng dữ liệu (NHDL) cho nhiều ngành trong nền kinh tế quốc dân. Ba mô hình NHDL phát triển mạnh đó là mô hình mạng lưới, mô hình phân cấp và mô hình dạng quan hệ. Đặc biệt mô hình dạng quan hệ có nhiều tiện lợi lớn cho người sử dụng và có độ độc lập dữ liệu cao.

Trong những năm gần đây việc áp dụng lý thuyết tập mờ vào xử lý thông tin đang phát triển mạnh mẽ. Trong bài này sẽ trình bày NHDL dạng quan hệ có áp dụng tập mờ nhằm xử lý các câu hỏi mờ và thông tin lưu trữ có các đặc trưng mờ.

1. Các khái niệm cơ bản:

Giả sử $U = \{ M_1, M_2, \ldots, M_n \}$ là một tập hợp hữu hạn các thuộc tính M_i, $i = 1, 2, \ldots, n$; mỗi thuộc tính có một miền giá trị tương ứng kí hiệu là dom (M_i). Một quan hệ R trên tập thuộc tính U là một tập hợp con của tích Đề các của các miền giá trị

$$R \subseteq \text{dom}\ (M_1) \times \text{dom}\ (M_2) \times \ldots \times \text{dom}\ (M_n).$$

Mỗi quan hệ đều được biểu diễn dưới dạng bảng. Mỗi cột tương ứng với một thuộc tính và mỗi hàng tương ứng với một bộ. Gọi t là một bộ thuộc R, ta có $t[M_i]$ là thành phần thứ i tương ứng thuộc tính M_i của bộ đó trong R.

Để khảo sát việc áp dụng tập mờ trong mô hình quan hệ, ta xem xét một vài ví dụ sau đây:

Trong thực tế thường bắt gặp các yêu cầu xử lý hay các mô tả một thực thể như: « tìm những người có lương cao », « một cô gái đẹp ». Các khái niệm « cao », « đẹp » thực chất là không rõ ràng. Để có thể mô tả các khái niệm không xác định như trên /ZADEH 1974/ và một số tác giả khác đã sử dụng hàm phân phối ngẫu nhiên f. Hàm f không xác định trên tập $\{0.1\}$ mà được xác định trên đoạn [0,1] dựa trên logic đa trị vô hạn. Hàm f được định nghĩa qua:

$$f_A : U \rightarrow [0,1]$$

và Zadeh gọi là hàm thực (membership function). Ta có định nghĩa tập mờ như sau;

Định nghĩa 1:

Tập mờ E trên tập U được đặc trưng bằng hàm f_E có độ phụ thuộc

$$f_E : U \rightarrow [0,1]$$

trong đó mỗi phần tử $u \in U$ được tương ứng 1 số thực $f_E(u)$ của đoạn [0,1). Giá trị $f_E(u)$ được Zadeh gọi là độ phụ thuộc mờ (gọi tắt là độ mờ).

Dựa trên cơ sở của tập mờ, dữ liệu trong NHDL dạng quan hệ có thể được phân chia làm 3 loại sau đây:

– Dữ liệu xác định, nếu chỉ một $u_i \in M$ sao cho

$$f_M(u) = \begin{cases} 1 & \text{với } u = u_i \\ 0 & \text{với } u \neq u_i \end{cases}$$

Trường hợp này cũng gọi là dữ liệu « rõ ».

– Dữ liệu đa trị, nếu như $f_M(u)$ là mờ.

– Dữ liệu không xác định, nếu

$f_M(u) = 1$ cho mọi $u \in M$ hoặc

$f_M(u) = 0$ cho mọi $u \in M$.

Trường hợp $f_M(u) = 1$ với mọi $u \in M$ là dữ liệu chưa biết (unknown)

$f_M(u) = 0$ với mọi $u \in M$ là dữ liệu không xác định (undefined) hoặc bằng không (null).

Ở đây chúng ta chỉ khảo sát các dạng dữ liệu nêu trên.

Trong /ZADEH 1971/ đã đưa ra khái niệm biến ngôn ngữ. Mỗi biến ngôn ngữ A đều có một tập đặc trưng mờ (fuzzy term) T(A). Ví dụ biến ngôn ngữ TUOI, ta có

T(tuổi) = {già trẻ, trung niên, rất trẻ,...}.

Các đặc trưng mờ được chia làm 2 loại là loại đơn giản như «già», «trẻ»,... và loại phức tạp như «rất già», «khoảng trung niên», ...; giá trị của các đặc trưng mờ loại đơn giản được xác định như sau:

$$f_{\text{trẻ}}(u) = \begin{cases} 1 \text{ với } u \leqslant 20 \\ \left(1 + \left(\dfrac{u-20}{5}\right)^2\right)^{-1} \text{ với } u > 20 \end{cases}$$

và loại phức tạp được biểu diễn qua loại đơn giản, ví dụ như:

$$f_{\text{rất trẻ}} = (f_{\text{trẻ}})^2, \ldots$$

Để phân biệt dữ liệu không xác định với các loại khác trong quá trình xử lý thông tin, chúng ta đưa thêm một tham số c. Tham số c cùng với độ phụ thuộc tạo thành một cặp $< f_E(u), c >$ hoặc đơn giản hơn là $< t, c >$. Theo giá trị P – chân lý cặp $< t, c >$ có nghĩa là một giá trị đúng hoặc một giá trị có thể, t có nghĩa như trong logic đa trị và nhận giá trị trong đoạn [0,1]. c là thành phần xác định và làm rõ cho t. c nhận giá trị T hoặc P, nếu c = P chứng tỏ t nhận giá trị không xác định (unknow, undefined, null), nếu c = T chứng tỏ t nhận giá trị đúng (xác định) hoặc giá trị mờ.

2. Hệ thống tìm kiếm mờ:

Để thuận tiện cho việc thiết kế chiến lược tìm kiếm mờ chúng ta sử dụng một số phép tính logic như sau:

Định nghĩa 2:

– Phép Conjunction

$$< t_1, c_1 > \wedge < t_2, c_2 > = < \min(t_1, t_2), \min(c_1, c_2) >$$

trong đó min (T, T) = T, min (T, P) = P, min (P, P) = P.

– Phép Disjunction

$$< t_1, c_1 > \vee < t_2, c_2 > = < \max(t_1, t_2), \max(c_1, c_2) >$$

trong đó max (T, T) = T, max (T, P) = T, max (P, P) = P.

– Phép phủ định (Negation)

$$\neg < t, T > = < 1 - t, T >$$

$$\neg < 1, P > = < 1, P >.$$

Chú ý rằng ở đây không định nghĩa cho $\neg < t, P >$ với $t \neq 1$ vì trong bài này chúng ta không dùng đến.

Định nghĩa 3 (hệ tìm kiếm mờ):

Một hệ tìm kiếm mờ được mô tả qua

RS = [R, T, Q ; Υ] với

R : quan hệ cơ sở nhận tất cả các bộ dữ liệu.

T : tập các đặc trưng mờ.

Q : tập tất cả các câu hỏi.

Υ : ánh xạ $Q \times R \rightarrow [0,1] \times \{0,1\}$.

Hàm Υ được định nghĩa đệ quy như sau:

S_1: Nếu câu hỏi là một thành phần d, trong đó có thể là một tân từ rõ hoặc mờ thì hàm của miền giá trị Ai xác định qua:

$$\Upsilon(d,x) = < f_d(u), c >$$

với $c \in \{T, P\}$, $u = x[Ai]$ là giá trị của miền Ai trong bộ x.

S_2: nếu q là một câu hỏi thì

$$\Upsilon(\neg q, x) = \neg \Upsilon(q, x)$$

$\neg(q,x)$ xác định qua định nghĩa 2.

S_3: Nếu p, q là 2 câu hỏi bất kỳ thì

$$\Upsilon(p \text{ AND } q, x) = \Upsilon(p,x) \wedge \Upsilon(q,x)$$

$$\Upsilon(p \text{ OR } q, x) = \Upsilon(p,x) \vee \Upsilon(q,x).$$

S_4: Mỗi câu hỏi $q \in Q$ đều được dẫn xuất qua việc áp dụng liên tiếp các bước từ S_1 đến S_3

3. Các phép tính về tập mờ :

Trong đoạn này chúng ta sẽ đưa ra một số định nghĩa về các phép tính trong tập mờ. Trên cơ sở đó chúng ta sẽ thiết lập một hệ thống trả lời Eq . Hệ thống trả trời Eq được định nghĩa như sau :

Định nghĩa 4 :

Gọi q là một câu hỏi. Hệ thống trả lời Eq là một tập mờ được xác định qua

$$f_{E_q}(x) = \bigcup_{x \in R} \Upsilon(q, x) \qquad \text{cho } \forall x \in R$$

hoặc có thể biểu diễn khác:

$$E_q = \bigcup_{x \in R} \Upsilon(q, x)/x \qquad \text{hoặc } E_q = \bigcup_{x \in R} (x ; \Upsilon(q, x)).$$

Các phép tính về tập hợp mờ:

Phép tính	Định nghĩa	Ghi chú
Phép hợp	$I_{E_p \cup E_q}(x) = f_{E_p}(x) \vee f_{E_q}(x) = \bigcup_{x_i \in R} (x_i ; < f_p(x_i), c_i > \vee < f_q(x_i), c'_i >),$ với $c_i, c'_i \in \{P, T\}, \forall x \in R$	
Phép giao	$f_{E_p \cap E_q}(x) = f_{E_p}(x) \wedge f_{E_q}(x) = \bigcup_{x_i \in R} (x_i ; < f_p(x_i), c_i > \wedge < f_q(x_i), c'_i >),$ Với $c_i, c'_i \in \{P, T\} \quad \forall x \in R$	
Tịch	$E_p . E_q = \bigcup_{x_i \in R} (x_i ; < f_p(x_i), c_{ip} > . < f_i(x_i), c_{iq} >) = \bigcup_{xi \in R} (x_i, < f_p(x_i) . i_q(x_i), c_{ip} . c_{iq} >$ với $c_{ip}, c_{iq} \in \{P, T\}$	P . P = P P . T = T . P = P T . T = T
Hàm mũ	$E_q^r = \bigcup_{x_i \in R} (x_i ; < f_q(x_i), c_i >^r) = \bigcup_{x_i \in R} (x_i ; < f_q^r(x_i), c_i >),$ r là số thực, $c_i \in \{P, T\}$	
Phép giãn Dilation	$DIL(E_q) = E_q^{0.5} = \bigcup_{x_i \in R} (x_i ; < f_q^{0.5}(x_i), c_i >)$ $c_i \in \{P, T\}$	Trường hợp đặc biệt của hàm mũ với r = 0.5
Phép nén Conzentration	$CON(E_q) = E_q^2 = \bigcup_{x_i \in R} (x_i ; < f_q^2(x_i), c_i >)$ $c_i \in \{P, T\}$	

Dựa vào các phép tính ở bảng có thể thiết lập một số chương trình con ngữ nghĩa để phục vụ cho việc tính toán giá trị các đặc trưng mờ phức tạp chẳng hạn các giá trị

Rất x = CON (Ex)
tương đối x = DIL (Ex)
Không rất x = NOTCON (Ex)

Trong đó x là một đặc trưng mờ đơn giản (ví dụ như già, trẻ, trung niên,...) và Ex là tập mờ của đặc trưng đó (xem / ZADEH 1974 /, / UMANO 1978 /).

4. Cấu trúc bộ nhớ:

Giả sử rằng tổ chức bộ nhớ của từng quan hệ R trong NHDL được tổ chức theo cấu trúc file ngược (xem / HISAO 1970 /, / HARDER 1978 /). Cấu trúc này cho phép xử lý nhanh để tìm các bộ thỏa mãn các yêu cầu xử lý. Nhưng sử dụng cấu trúc này tương đối tốn bộ nhớ. Ở đây các quan hệ không được lấy file ngược hoàn toàn. Các thuộc tính thuộc quan hệ sẽ được lấy file ngược cần phải thỏa mãn một số tiêu chuẩn nêu ở phần sau. Trong bài này chúng ta sẽ sử dụng khái niệm TID (Tuple Identifier) và miền giá trị liên hệ (domain related). M_i và M_j là hai thuộc tính.

$dom(M_j)$ và $dom(M_j)$ là 2 miền giá trị rõ (xác định) tương ứng của 2 thuộc tính M_i và M_j được gọi là 2 thuộc tính có miền giá trị liên hệ nếu

$$dom(M_i) \cap dom(M_j) \neq \phi \ , \ i \neq j.$$

Ta ký hiệu là $M_i \sim M_j$.

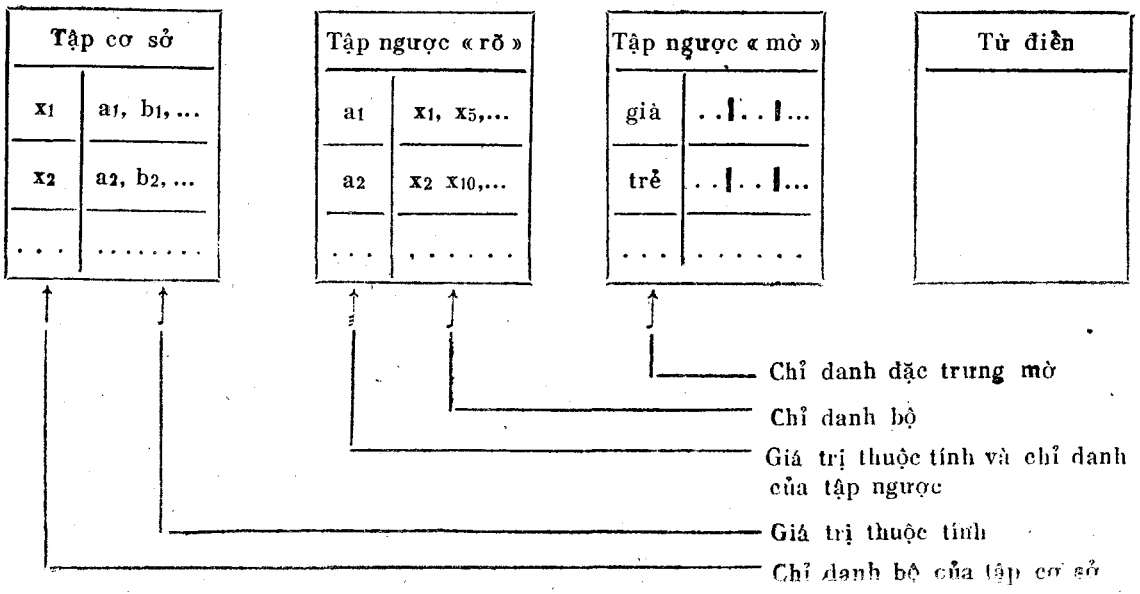
Trên cơ sở giả thiết và kết quả nêu ở các phần trên, dữ liệu trong NHDL có thể được phân chia và tổ chức như sau:

a/ Tập gốc bao gồm tất cả các bộ của một quan hệ, trong đó mỗi bộ có một địa chỉ ô nhớ. Ta ký hiệu là x_i

b) Mỗi quan hệ có một tập ngược « rõ ». Tập này bao gồm một tập danh sách chỉ số. Cho mỗi giá trị thuộc tính có một danh sách chỉ số tương ứng.

Các thuộc tính được lấy tập ngược phải thỏa mãn một trong các điều kiện sau đây:

+ Các thuộc tính có miền giá trị liên hệ.
+ Khóa chính,
+ Các thuộc tính có tần suất khai thác cao.



Hình 1 – Tổ chức của quan hệ chia làm 4 phần

Biểu diễn logic 1 danh sách tập ngược « rõ » của một quan hệ như sau:

Khóa	Tên quan hệ	Tên thuộc tính	Số lượng bộ	TID	TID	...	...	TID

Tên thuộc tính	Số lượng bộ	TID	...	...		TID

c) Để xử lý các câu hỏi mờ ta tổ chức tập ngược cho tất cả các đặc trưng mờ đơn giản có thể xảy ra trong NHDL. Tập ngược mờ bao gồm một tập danh sách chỉ số mờ. Mỗi danh sách chỉ số mờ bao gồm 3 phần:

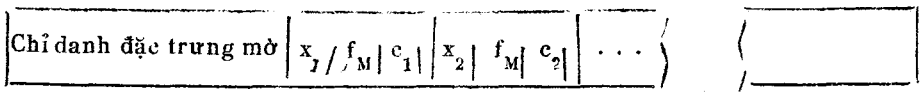
– Địa chỉ các bộ của quan hệ cơ sở.

– Độ phụ thuộc của tất cả các giá trị của thuộc tính mà tương ứng với đặc trưng mờ này. Việc tính toán độ phụ thuộc sẽ được các chương trình con ngữ nghĩa đảm nhiệm.

– Một số thông tin đặc biệt nhằm phân biệt dữ liệu xác định và không xác định.

d) Từ điển lưu giữ tất cả các đặc trưng mờ đơn giản và phức tạp và các chương trình con ngữ nghĩa để giúp cho việc thiết lập các tập mờ.

Cấu trúc của danh sách chỉ số mờ:



5. Chiến lược tìm kiếm:

Dựa trên các kết quả trên và cấu trúc bộ nhớ cho NHDL chúng ta có một số chiến lược tìm kiếm được thể hiện ở các định lý sau đây:

Định lý 1. Giả sử q là một câu hỏi thuộc tập câu hỏi Q. Nếu q là một tân từ $M_j = d$, trong đó M_j là một thuộc tính thì hệ thống trả lời E_q được xác định như sau:

$$E_q = \bigcup_{u \in dom(M_j)} (IL(u); < f_d(u), c_u >),$$

trong đó $(IL(u); < f_d(u), c_u >)$ là một tập mờ và $c_u \in \{P, T\}$ cho giá trị u, IL(u) là danh sách chỉ số tập ngược «rõ» của giá trị u. Hàm f trên tập mờ được xác định qua

$$f_{(IL(u); < f_d(u), c_u >)}(x) = \begin{cases} \bigcup_{u \in R} < f_d(x), c_u > \text{ cho } x \in IL(u) \\ 0 \text{ trường hợp còn lại.} \end{cases}$$

Chứng minh: Nếu d là dữ kiện xác định hoặc không xác định như unknown, undefined, null thì việc kiểm tra không có gì khó khăn. Chúng ta chỉ kiểm tra cho d là một đặc trưng mờ.

(ta tạm ký hiệu q = «t»)

Cho mỗi $x \in R$ cần chỉ ra rằng

$$f_{E_q}(x) = f_{\bigcup_u (IL(u); < f_t(u), c_u >)}(x) \quad (1)$$

Theo định nghĩa 4 vế trái của (1) là $f_{E_q}(x) = \bigvee \gamma(q, x)$

Theo định nghĩa của hàm γ ta có:

$\gamma(q, x) = < f_t(u), c_u >$, trong đó $u = x[M_j]$. Từ đó ta có:

$$f_{E_q}(x) = \bigcup_{u \in R} < f_t(u), c_u > \quad (2)$$

Theo định nghĩa trong bảng 1 vế phải của (1) là

$$f_{\bigcup_{u \in R} (IL(u); < f_t(u), c_u >)}(x) = f_{(IL(u), < f_t(u), c_u >)}(x) \vee$$

$$f_{(IL(u'); < f_t(u'), c_{u'} >)}(x) \vee \ldots$$

trong đó $(IL(v); < f_t(v), c_v >)$ là tập mờ trong R và

$$f_{(IL(v); < f_t(v), c_v >)}(x) = \begin{cases} \bigcup_{v \in R} < f_t(v), c_v > \text{ cho } x \in TL(v), c_v \in \{p, T\}. \\ 0 \text{ còn lại} \end{cases}$$

Vì rằng $\underline{u} = x[M_i]$ và $\underline{u}$ chỉ là một giá trị của dom (M_i) trong bộ x, cho nên $x \in IL(u)$. Nhưng $x \in IL(u')$ và $x \in IL(u'')$, v.v... Từ đó ta có:

$$f_{\cup}(IL(u); \langle f_t(u), c_u \rangle)(x) = \bigcup_{u \in R} \langle f_t(u), c_u \rangle \quad (3)$$

Kết hợp (2) và (3) ta có điều cần chứng minh.

Định lý 2. Cho 2 câu hỏi p và q bất kỳ thuộc tập Q. Ta luôn luôn có:

$$E_{pANDq} = E_p \cap E_q \quad (4)$$

và

$$E_{pORq} = E_p \cup E_q; \quad (5)$$

trong đó $\cap$ và $\cup$ là 2 phép tính giao, hợp của tập mờ.

Mô hình quan hệ có áp dụng xử lý tập mờ và câu hỏi mờ trên đây được thiết kế và cài đặt thử nghiệm trên máy EC 1020. Để xử lý các câu hỏi mờ, mô hình được cài đặt thêm một Buildin – Function (hàm mẫu) Fuzzy. Các chương trình và các phép tính quan hệ xử lý tập mờ được viết bằng ngôn ngữ PL/1.

Nhận ngày 15-12-1984

TÀI LIỆU THAM KHẢO

1. **D.D.Chamberlin, M.M.Astrahan...**, SEQUEL-2; A unified Approach to Data Definition, Manipulation and Control, IBM J. Res, Dev. 20, 6(1978) 560–575.

2. **E F. Codd**, A relational Model of Data for large Schared Data Banks C.ACM 13, 6 (1970)

3. **Th. Haerder**, Implementierung von Datenbankssystemen, Munchen, Wien, Carl Hauer–Verlag 1978.

4. **D.Hsiao, F.Farary**, A formal System foı information retrievat From files, C. ACM 13, 2(1970).

5. **V. Tahani**, A fuzzy model of document retrieval System, Inf. Proc. Man. 12 (1976).

6. **V. Tahani**, A conceptual framework foı fuzzy query processing – a step toward very intelligent database system, Inf. Pro. Man. 13 (1977).

7. **M. Umano**, Respresentation and Manipulation of fuzzy data, Diss. Osaka Uni. Japan 2.1979.

8. **Le Tien Vuong**, Untersuchung zur ternaeren Dekomposition einer Relation und zur Anwendung der unscharfen Mengen im CRM, Diss. TU-Dresden 1983.

9. **L.A.Zadeh**, Fuzzy logic and its application to approximate reasoning, Inf. Pro. 1974

ABSTRACT

On the application of fuzzy set theory in relational database

In this paper we introduced a model of retrieval system based on the theory of fuzzy sets in which we constructed a suitable storage structure with incomplete inverted files (they contain even some incomplete inverted files for a collection of fuzzy data of database). On this concept, we introduced a method for fuzzy query processing in the system and shown some retrieval strategies.

The general response system was also defined and the ability of these strategies was investigated.