Một mô hình cơ sở dữ liệu quan hệ mờ

A Fuzzy Relational Data Base Model

Nguyễn Hòa

Abstract: *This paper introduces a fuzzy relational data base model (FRDB) that extends the conventional relational data base model with two key features: (1) the relations represent the set of data tuples to be the fuzzy relations; (2) selection conditions are associated with fuzzy set values to be able to query the fuzzy, imprecise information of objects in relations. An interpretation of the membership degree of tuples for fuzzy relations is proposed on the basis of the fuzzy set theory to develop the data model and data manipulating model of FRDB that consist of schemas, fuzzy relations and algebraic operations. Some properties of the fuzzy relational algebraic operations are also formulated and proven as extensions of the properties of relational algebraic operations in the conventional relational data base model.*

Keywords*: Fuzzy set, fuzzy relation, fuzzy relational data base, fuzzy relational algebraic operation.*

I. GIỚI THIỆU

Như chúng ta đã biết, mô hình *cơ sở dữ liệu quan hệ truyền thống* (conventional relational data base), được đề nghị bởi Codd E.F năm 1970 [2], đã chứng tỏ nhiều ưu điểm trong các vấn đề mô hình hóa, thiết kế và hiện thực các hệ thống lớn, từ phần mềm cho đến cơ sở dữ liệu (CSDL). Tuy nhiên, các ứng dụng dựa trên mô hình CSDL quan hệ truyền thống không biểu diễn được các đối tượng mà thông tin về chúng không được xác định một cách rõ ràng và chính xác. Điều đó làm hạn chế khả năng mô hình hóa và giải quyết các bài toán áp dụng trong thế giới thực. Chẳng hạn, các ứng dụng mô hình CSDL truyền thống không thể trả lời được các truy vấn kiểu như "tìm tất cả bệnh nhân *trẻ* có tiền sử bệnh viêm thanh quản"; hoặc "tìm tất cả các gói bưu kiện có trọng lượng *khoảng 10* kg và được vận chuyển trong thời gian *khoảng 36* giờ từ Hà Nội đến Sài Gòn", v.v. Trong đó *trẻ, khoảng 10* kg và *khoảng 36* giờ là những khái niệm và giá trị không chính xác.

Trong những năm qua đã có nhiều mô hình cơ sở dữ liệu quan hệ dựa trên lý thuyết *tập mờ* (fuzzy set) được nghiên cứu và xây dựng [1,3,7,9,10], nhằm mô hình hóa các đối tượng mà thông tin về chúng mờ, không chính xác để khắc phục các hạn chế của mô hình CSDL quan hệ truyền thống. Các mô hình như vậy gọi là *mô hình cơ sở dữ liệu quan hệ mờ* (fuzzy relational data base model).

Có hai cách tiếp cận chính để biểu diễn dữ liệu mờ trong mô hình CSDL mờ: (1) biểu diễn giá trị thuộc tính bằng các giá trị tập mờ trong quan hệ mờ; (2) biểu diễn giá trị thuộc tính bằng các giá trị đơn, chính xác trong quan hệ mờ.

Trong cách tiếp cận thứ nhất, giá trị thuộc tính quan hệ được biểu diễn bằng một tập mờ và được diễn dịch bởi hàm thành viên của nó [4,7,9,11]. Trong các mô hình được xây dựng bằng cách tiếp cận này, các quan hệ hai ngôi cổ điển giữa các thuộc tính được mở rộng thành các quan hệ mờ. Mức độ thành viên (độ thuộc) của các bộ được ẩn trong mức độ thành viên của các giá trị thuộc tính. Cách tiếp cận thứ hai thực chất là mở rộng các quan hệ cổ điển nhiều ngôi thành các quan hệ mờ nhiều ngôi trên các miền giá trị thuộc tính quan hệ. Với các mô hình được xây dựng theo các tiếp cận này, mức độ thành viên của giá trị thuộc tính được ẩn trong mức độ thành viên của các bộ trong các quan hệ mờ.

Mặc dù có một số nghiên cứu phát triển CSDL quan hệ mờ dựa trên cách tiếp cận thứ hai [6,10],

nhưng theo chúng tôi được biết, chưa có một mô hình nào được xây dựng bằng cách định nghĩa các khái niệm biểu diễn dữ liệu và các phép toán đại số quan hệ một cách hình thức để từ đó phát biểu và chứng minh các tính chất của các phép toán đại số cũng như xây dựng một hệ quản trị CSDL cho mô hình.

Bài báo này đề nghị một mô hình CSDL quan hệ mờ (FRDB) dựa trên cách tiếp cận thứ hai, trong đó các khái niệm lược đồ, quan hệ và các phép toán đại số quan hệ được định nghĩa một cách hình thức như những mở rộng trực tiếp các khái niệm và phép toán tương ứng trong CSDL truyền thống. Đặc biệt, các điều kiện chọn được phân loại và ngữ nghĩa của chúng được diễn dịch như một ánh xạ cho phép kết hợp với các tập mờ để thực hiện các *truy vấn mềm* (soft query) trên các quan hệ mờ trong FRDB. Chúng tôi cũng phát biểu và chứng minh một số tính chất của các phép toán đại số trên FRDB và phát triển một hệ quản trị khởi đầu cho FRDB để minh họa cho triển vọng ứng dụng của mô hình này vào thực tế.

Cơ sở toán học để phát triển FRDB được trình bày trong Phần II, lược đồ và thể hiện FRDB được giới thiệu trong Phần III. Phần IV trình bày các phép toán đại số trên FRDB, Phần V giới thiệu một hệ quản trị khởi đầu của FRDB và cuối cùng Phần VI là một số kết luận và hướng nghiên cứu trong tương lai.

II. CƠ SỞ TOÁN HỌC CỦA FRDB

Phần này giới thiệu tập mờ và quan hệ mờ [6] như là cơ sở toán để phát triển FRDB. Tập mờ được sử dụng để biểu diễn các truy vấn với thông tin không rõ ràng, quan hệ mờ được sử dụng để định nghĩa các quan hệ trong FRDB.

II.1. Tập mờ

Tập mờ là khái niệm mở rộng của tập cổ điển và được định nghĩa như sau.

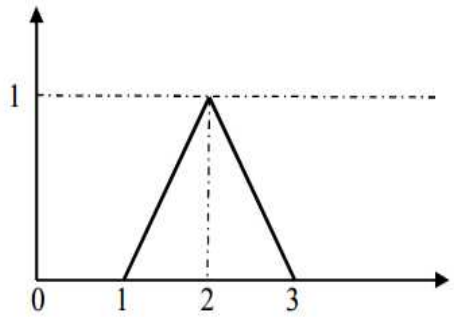
Định nghĩa 1. Giả sử $\mathcal{X}$ là một tập khác rỗng, một ánh xạ từ $\mathcal{X}$ đến khoảng đóng [0, 1], $\mu_A: \mathcal{X} \rightarrow [0, 1]$, xác định *một tập mờ* (fuzzy set) *A* trên $\mathcal{X}$. Ánh xạ μ_A được gọi là *hàm thành viên* (membership function) của tập mờ *A*. Với mỗi $x \in \mathcal{X}$, $\mu_A(x)$ là *mức độ thành viên* (membership degree) của *x* đối với *A*.

Để đơn giản, ký hiệu $A: \mathcal{X} \rightarrow [0, 1]$ có thể được sử dụng để biểu diễn tập mờ *A*.

Ví dụ 1. Một ví dụ đơn giản về tập mờ là tập các số gần số 2, *about_2*, được cho bởi hàm thành viên của nó như sau:

$$about_2 = \begin{cases} x-1 \ \forall x \in [1,2] \\ 3-x \ \forall x \in (2,3] \\ 0 \ \forall x \notin [1,3] \end{cases}$$

và đồ thị hàm thành viên của *about_2* như trong Hình 1.



Hình 1. Tập mờ các số gần 2

Các phép toán trên các tập mờ được định nghĩa một cách tổng quát dựa trên các ánh xạ từ tập tích Descartes của các khoảng đóng [0,1] đến khoảng đóng [0,1]. Tuy nhiên, phần này chỉ giới thiệu các *phép toán chuẩn* (standard operation) trên các tập mờ [6] được ứng dụng trong FRDB.

Định nghĩa 2. Giả sử *A*, *B* là hai tập mờ trên tập $\mathcal{X}$ và có các hàm thành viên lần lượt là μ_A, μ_B. Phép toán lấy phần bù của *A*, hợp, giao và hiệu của *A* và *B* được định nghĩa theo hàm thành viên của chúng như sau.

1. $\mu_{A^c}(x) = 1-\mu_A(x), \forall x \in \mathcal{X}$
2. $\mu_{A\cup B}(x) = max(\mu_A(x), \mu_B(x)), \forall x \in \mathcal{X}$
3. $\mu_{A\cap B}(x) = min(\mu_A(x), \mu_B(x)), \forall x \in \mathcal{X}$
4. $\mu_{A-B}(x) = min(\mu_A(x), 1-\mu_B(x)), \forall x \in \mathcal{X}$.

II.2. Quan hệ mờ

Quan hệ mờ được định nghĩa bằng cách mở rộng quan hệ cổ điển như sau.

Định nghĩa 3. Giả sử $A_1, A_2,\ldots, A_k$ là các tập khác rỗng, một quan hệ mờ k-ngôi R giữa k tập $A_1, A_2,\ldots, A_k$ là một tập con mờ của tập tích Descartes $A_1\times A_2 \times\ldots\times A_k$.

Như vậy, một quan hệ mờ k-ngôi R được kết hợp với một hàm thành viên μ_R: $A_1\times A_2\times\ldots\times A_k \rightarrow[0,1]$.

III. LƯỢC ĐỒ VÀ QUAN HỆ FRDB

Lược đồ và quan hệ FRDB được mở rộng từ lược đồ và quan hệ CSDL truyền thống để biểu diễn khả năng các bộ thuộc về một quan hệ mờ.

III.1. Lược đồ FRDB

Một lược đồ FRDB gồm một tập thuộc tính kết hợp với một hàm thành viên, làm cơ sở để xác định các quan hệ mờ, được định nghĩa như sau.

Định nghĩa 4. Một *lược đồ quan hệ mờ* (fuzzy relational schema) là một bộ đôi $R = (\boldsymbol{U}, \mu)$, trong đó

1. $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ là một tập các thuộc tính đôi một khác nhau (biểu diễn thông tin về giá trị các đối tượng trong quan hệ).
2. μ là một ánh xạ đặt tương ứng mỗi $(v_1, v_2, \ldots, v_k) \in D_1\times D_2\times\ldots\times D_k$ với một số thực thuộc [0, 1], trong đó D_i là miền giá trị của thuộc tính A_i (i = 1, …, k).

Chúng tôi lưu ý rằng, như trong CSDL quan hệ truyền thống, để đơn giản, có thể viết $R(\boldsymbol{U}, \mu)$ thay cho cách viết $R = (\boldsymbol{U}, \mu)$. Ngoài ra, mỗi $t = (v_1, v_2, \ldots, v_k)$ được gọi là một bộ trên tập thuộc tính $\{A_1, A_2, \ldots, A_k\}$.

Ví dụ 2. Một lược đồ quan hệ mờ **PATIENT** trong FRDB mô tả về các bệnh nhân có thể như sau:

PATIENT(PATIENT_ID, PATIENT_NAME, AGE, SEX, μ), với μ là ánh xạ

μ: **string** × **string** ×**real** × **binary**→[0, 1], trong đó **string**, **real** và **binary** là các miền giá trị của các thuộc tính PATIENT_ID, PATIENT_NAME, AGE và SEX.

III.2. Quan hệ FRDB

Quan hệ mờ được mở rộng từ quan hệ truyền thống [2, 5] với mức độ thuộc được gán cho mỗi bộ như định nghĩa dưới đây.

Định nghĩa 5. Giả sử $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ là một tập thuộc tính đôi một khác nhau, một *quan hệ mờ* (fuzzy relation) r trên lược đồ $R(\boldsymbol{U}, \mu)$ là một tập hữu hạn các bộ $\{t_1, t_2,\ldots, t_n\}$ trên tập các thuộc tính $\{A_1, A_2, \ldots, A_k\}$, được kết hợp tương ứng với các giá trị $\mu(t_i)$ biểu diễn mức độ thuộc của t_i trong r. Các ký hiệu $t.A$ hoặc $t[A]$ biểu thị giá trị thuộc tính A của bộ t trong r. Mức độ thuộc của t_i trong r được ký hiệu là $\mu_r(t_i)$.

Với mỗi tập thuộc tính $\boldsymbol{X} \subseteq \{A_1, A_2, \ldots, A_k\}$, ký hiệu $t[\boldsymbol{X}]$ được sử dụng để biểu thị phần còn lại của t sau khi loại bỏ các giá trị của các thuộc tính không thuộc X.

Phụ thuộc hàm mờ (fuzzy function dependence) trong FRDB được mở rộng từ phụ thuộc hàm trong CSDL truyền thống [10].

Định nghĩa 6. Giả sử $R(\boldsymbol{U}, \mu)$ là một lược đồ quan hệ mờ, r là một quan hệ mờ bất kì trên $R(\boldsymbol{U}, \mu)$, X và Y là hai tập con các thuộc tính của $\boldsymbol{U}$. Ta nói Y là *phụ thuộc hàm mờ* đối với X trên lược đồ $R(\boldsymbol{U}, \mu)$ và ký hiệu $X \rightsquigarrow Y$, nếu

$\forall t_1, t_2 \in r, (\mu_r(t_1) \wedge \mu_r(t_2) \wedge (t_1[X] = t_2[X])) \rightarrow (t_1[Y] = t_2[Y])$, trong đó $\wedge$ là một phép toán hội logic mờ được định nghĩa bởi $a \wedge b = min(a, b)$ (do Zadeh đề nghị trong [6]) và suy luận mờ $a \rightarrow b$ được định nghĩa bởi $a \rightarrow b = 1$ nếu $a \leq b$, và $a \rightarrow b = 1 - (a-b)$ nếu ngược lại, với $a, b \in [0, 1]$ (do Lukasiewicz đề nghị trong [6]). Suy luận mờ này là một mở rộng trực tiếp từ suy luận trong logic cổ điển ($a \rightarrow b = 1$ nếu $a \leq b$ và $a \rightarrow b = 0$, nếu ngược lại, với $a, b \in \{0, 1\}$).

Như vậy, phụ thuộc hàm mờ là phụ thuộc theo *mức độ* (gradual), nghĩa là phụ thuộc càng lớn khi giá trị của suy luận mờ càng lớn.

Chúng tôi lưu ý rằng, khi giá trị hàm μ_r bằng 1 với mọi r trên $R(\boldsymbol{U}, \mu)$, suy luận mờ xác định phụ thuộc hàm lúc đó đồng nhất với suy luận trong logic cổ điển (do các mệnh đề tham gia $(t_1[X] = t_2[X])$ và $(t_1[Y] = t_2[Y])$ chỉ có thể nhận các giá trị 0 hoặc 1) và phụ

thuộc hàm mờ trong định nghĩa này sẽ đồng nhất với phụ thuộc hàm trong CSDL truyền thống [5].

Bây giờ CSDL quan hệ mờ là mở rộng của CSDL quan hệ truyền thống và được định nghĩa như sau.

Định nghĩa 7. Một *cơ sở dữ liệu quan hệ mờ* (fuzzy relational database) trên một tập các thuộc tính $\mathcal{A}$ là một tập các quan hệ mờ tương ứng với tập các lược đồ quan hệ mờ của chúng.

Lưu ý rằng, nếu chỉ quan tâm đến một quan hệ duy nhất trên một lược đồ thì có thể đồng nhất ký hiệu tên quan hệ và lược đồ của chúng.

Ví dụ 3. Một CSDL quan hệ mờ đơn giản các bệnh nhân tại phòng khám của một bệnh viện có thể được tổ chức như các Bảng 1, 2 và 3. Trong đó PATIENT (BỆNH NHÂN), DIAGNOSE (KHÁM BỆNH) và PHYSICIAN (BÁC SĨ) là các quan hệ mờ trên các lược đồ cùng tên. Các thuộc tính PATIENT_ID, PATIENT_NAME, AGE, WEIGHT, MEDICAL_ HISTORY, DISEASE, DURATION, COST tương ứng biểu diễn mã số, tên, tuổi, trọng lượng, tiền sử bệnh, bệnh (được chẩn đoán hiện tại), thời gian điều trị và chi phí điều trị trên ngày của một bệnh nhân.

Các thuộc tính khác như PHYSICIAN_ID, PHYSICIAN_NAME và EXPERIENCE tương ứng biểu diễn mã số, tên và thâm niên khám chữa bệnh của một bác sĩ. Một số bệnh tiền sử hoặc được chẩn đoán hiện tại của các bệnh nhân trong CSDL như *Bronchitis* (viêm phế quản), *Gall-stone* (sỏi mật), *Hepatitis* (viêm gan), *Tuberculosis* (lao phổi), *Lung cancer* (ung thư phổi). Ở đây, quy ước đơn vị thời gian điều trị, chi phí điều trị tương ứng là *ngày* và 1000 (đồng VN). Thâm niên khám chữa bệnh của bác sĩ được tính theo năm.

Chúng tôi lưu ý rằng, giá trị $\mu(t)$ biểu diễn độ thuộc của mỗi bộ t trong các quan hệ mờ, khi $\mu(t) = 1.0$, bộ t thực sự thuộc quan hệ. Xét về toán học, ý nghĩa độ thuộc $\mu(t)$ của t (Định nghĩa 5) trong quan hệ mờ tương tự như độ thuộc của một phần tử trong tập mờ (quan hệ mờ cũng là một tập mờ). Xét về khả năng biểu diễn trong ứng dụng, độ thuộc cho biết mức độ chính xác của thông tin. Chẳng hạn, xét bộ t_1 (bộ thứ nhất) trong quan hệ DIAGNOSE, giả sử thông tin về mã số bệnh nhân (*PT005*) và mã số bác sĩ (*DT001*) là chính xác, lúc này độ thuộc $\mu(t_1) = 0.7$ của t_1 biểu diễn độ chính xác của bệnh được chẩn đoán (*Tuberculosis*-lao phổi), thời gian điều trị (400 ngày) và chi phí điều trị (300 ngàn đồng/ngày) của bệnh nhân.

Chúng tôi lưu ý thêm rằng, CSDL mờ ở đây là đơn giản, các giá trị thuộc tính cũng như việc gán độ thuộc cho các bộ của các quan hệ chỉ có ý nghĩa minh họa cho các Định nghĩa 4, 5 và 7. Trong các ứng dụng thực tế, xác định độ thuộc của các bộ trong một quan hệ mờ là một phần của quá trình thiết kế CSDL và được các chuyên gia thiết kế, chuyên gia lĩnh vực xác định dựa trên việc đánh giá độ chính xác của các giá trị thuộc tính của các bộ trong quan hệ.

Bảng 1. Quan hệ PATIENT

PATIENT_ID	PATIENT_NAME	AGE	WEIGHT	MEDICAL_HISTORY	μ
PT005	*L.V. Tam*	53	70	*Bronchitis*	0.9
PT006	*N..T. Trang*	29	49	*Gall-stone*	0.5
PT007	*T. T. Tu*	21	65	*Hepatitis*	1.0

Bảng 2. Quan hệ PHYSICIAN

PHYSICIAN_ID	PHYSICIAN_NAME	EXPERIENCE	μ
DT001	*N. T. Son*	30	0.6
DT002	*H. V. Tuan*	25	0.8
DT003	*T. T. T. Nhan*	6	0.9

Bảng 3. Quan hệ DIAGNOSE

PATIENT_ID	PHYSICIAN_ID	DISEASE	DURATION	COST	μ
PT005	*DT001*	*Tuberculosis*	400	300	0.7
PT006	*DT002*	*Hepatitis*	40	30	0.5
PT007	*DT003*	*Lung cancer*	500	350	0.4

IV. CÁC PHÉP TOÁN ĐẠI SỐ FRDB

Các phép toán đại số quan hệ mờ như phép chọn, phép giao, phép hợp và phép trừ là cơ sở để truy vấn và thao tác dữ liệu mờ, không chính xác trong FRDB. Các phép toán này được mở rộng từ các phép toán đại số quan hệ truyền thống, trong đó mức độ thành viên của các bộ là một giá trị trong khoảng [0, 1].

IV.1. Phép chọn

Phép chọn trên một quan hệ FRDB là cơ sở để thực hiện các truy vấn tìm kiếm thông tin trong CSDL. Trước khi định nghĩa phép chọn, chúng tôi giới thiệu cú pháp và ngữ nghĩa của các điều kiện chọn như dưới đây.

Định nghĩa 8. Giả sử R là một lược đồ FRDB, χ là một tập các biến bộ quan hệ, θ là một quan hệ hai ngôi trong {=, ≠, ≤, <, >, ≥}. Các *điều kiện chọn mờ* (fuzzy selection condition) được định nghĩa một cách đệ quy và có một trong các dạng sau:

1. $x.A\ \theta\ v$, trong đó $x \in \chi$, A là một thuộc tính trong R và v là một giá trị.
2. $x.A \rightarrow v$, trong đó $x \in \chi$, A là một thuộc tính trong R, $\rightarrow$ là một quan hệ hai ngôi mờ và v là một *giá trị tập mờ* (fuzzy set value).
3. $x.A_1\ \theta\ x.A_2$, trong đó $x \in \chi$, A_1 và A_2 là hai thuộc tính phân biệt trong R.
4. $\neg E$ nếu E là một điều kiện chọn mờ.
5. $E_1 \wedge E_2$ nếu E_1 và E_2 là các điều kiện chọn mờ trên cùng một biến bộ quan hệ.
6. $E_1 \vee E_2$ nếu E_1 và E_2 là các điều kiện chọn mờ trên cùng một biến bộ quan hệ.

Chúng tôi lưu ý rằng, quan hệ hai ngôi mờ $\rightarrow$ trong $x.A \rightarrow v$ với v là một tập mờ biểu diễn truy vấn "tìm tất cả các bộ t sao cho $t.A$ là v", quan hệ hai ngôi $\theta \in$ {=, ≠, ≤, <, >, ≥} trong $x.A\ \theta\ v$ và $x.A_1\ \theta\ x.A_2$ có ý nghĩa tương tự như trong CSDL truyền thống.

Ví dụ 4. Với lược đồ quan hệ **PATIENT** trong CSDL các bệnh nhân ở Ví dụ 3, một số điều kiện chọn mờ có thể như sau (x là biến bộ):

1. Tìm tất cả bệnh nhân *trẻ tuổi* (*young*) và có tiền sử bệnh viêm gan (*hepatitis*). Yêu cầu này có thể được biểu diễn bởi điều kiện chọn x.AGE $\rightarrow$ *young* $\wedge x$.MEDICAL_HISTORY=*hepatitis*.
2. Tìm tất cả bệnh nhân *cao tuổi (old)* hoặc có cân nặng dưới 50 kg. Yêu cầu này có thể được biểu diễn bởi điều kiện chọn x.AGE $\rightarrow$ *old* $\vee$ x.WEIGHT < 50.

Định nghĩa 9. Giả sử $R(\boldsymbol{U}, \mu)$ là một lược đồ quan hệ FRDB, r là một quan hệ trên R, x là một biến bộ quan hệ và t là một bộ trong r. *Diễn dịch* (interpretation) của các điều kiện chọn mờ theo R, r và t, được biểu thị bởi $int_{\boldsymbol{R},r,t}$, là một ánh xạ bộ phận từ tập tất cả các điều kiện chọn mờ đến khoảng [0, 1] và được định nghĩa đệ qui như sau:

1. $int_{\boldsymbol{R},r,t}(x.A\ \theta\ v) = \mu_r(t)$ nếu $t.A\ \theta\ v$ và $int_{\boldsymbol{R},r,t}(x.A\ \theta\ v) = 0$ nếu ngược lại.
2. $int_{\boldsymbol{R},r,t}(x.A \rightarrow v) = min(\mu_r(t), \mu_\varphi(t))$, với $\varphi = t.A \rightarrow v$.
3. $int_{\boldsymbol{R},r,t}(x.A_1\ \theta\ x.A_2) = \mu_r(t)$ nếu $t.A_1\ \theta\ t.A_2$ và $int_{\boldsymbol{R},r,t}(x.A_1\ \theta\ x.A_2) = 0$ nếu ngược lại.
4. $int_{\boldsymbol{R},r,t}(\neg E) = 1 - int_{\boldsymbol{R},r,t}(E)$
5. $int_{\boldsymbol{R},r,t}(E_1 \wedge E_2) = min(int_{\boldsymbol{R},r,t}(E_1), int_{\boldsymbol{R},r,t}(E_2))$
6. $int_{\boldsymbol{R},r,t}(E_1 \vee E_2) = max(int_{\boldsymbol{R},r,t}(E_1), int_{\boldsymbol{R},r,t}(E_2))$.

Chúng tôi lưu ý rằng, v là một tập mờ trong $t.A \rightarrow v$ nên $\varphi = t.A \rightarrow v$ là một quan hệ mờ. Vì vậy φ cũng là một tập mờ. Cụ thể φ là tập mờ mà hàm thành viên của nó có đối số là các bộ t của r. Với mỗi $t \in r$, $\mu_\varphi(t) = v(t.A)$.

Về trực giác, $int_{\boldsymbol{R},r,t}(x.A\ \theta\ v)$ và $int_{\boldsymbol{R},r,t}(x.A \rightarrow v)$ tương ứng cho biết mức độ thỏa mãn các điều kiện $t.A\ \theta\ v$ và $t.A \rightarrow v$ (ở đây v là tập mờ) của bộ t trong r còn $int_{\boldsymbol{R},r,t}(x.A_1\ \theta\ x.A_2)$ cho biết mức độ thỏa mãn điều kiện $t.A_1\ \theta\ t.A_2$ của bộ t trong r. Ngoài ra, do giá trị v và các giá trị của các thuộc tính A, A_1 và A_2 trong $x.A\ \theta\ v$ và $x.A_1\ \theta\ x.A_2$ là các giá trị rõ, chính xác (Định nghĩa 4 và 5), nên nếu quan hệ θ không thỏa mãn trong $t.A\ \theta\ v$ và $t.A_1\ \theta\ t.A_2$ thì $int_{\boldsymbol{R},r,t}(x.A\ \theta\ v) = 0$ và $int_{\boldsymbol{R},r,t}(x.A_1\ \theta\ x.A_2) = 0$, nghĩa là bộ t tương ứng hoàn toàn không thỏa mãn $t.A\ \theta\ v$ và $t.A_1\ \theta\ t.A_2$.

Bảng 4. Quan hệ r'= σ_ϕ(PATIENT)

PATIENT_ID	PATIENT_NAME	AGE	WEIGHT	MEDICAL_HISTORY	μ
PT007	*T. T. Tu*	21	65	*Hepatitis*	0.93

Trong CSDL truyền thống vì $\mu_r(t) \in \{0, 1\}$ nên diễn dịch các điều kiện chọn đối với một bộ *t* luôn chỉ nhận một trong hai giá trị là 0 hoặc 1 (giá trị của $\mu_r(t)$). Vì vậy, có thể coi diễn dịch các điều kiện chọn trong CSDL truyền thống là trường hợp riêng của diễn dịch các điều kiện chọn trong FRDB.

Ví dụ 5. Giả sử tập mờ *young* biểu diễn tuổi trẻ của một bệnh nhân với hàm thành viên được định nghĩa bởi

$$young = \begin{cases} 1 \ \forall x \in [0, 20] \\ (35 - x)/15 \ \forall x \in (20, 35), \\ 0 \ \forall x \geq 35 \end{cases}$$

thì diễn dịch của điều kiện chọn mờ *E* = "*x*.AGE → *young* ∧ *x*.MEDICAL_HISTORY = *hepatitis*" theo quan hệ *r* = PATIENT và bộ t_3 (bộ thứ 3) trong CSDL các bệnh nhân ở Ví dụ 3 là $int_{\boldsymbol{R},r,t3}(E) = min(min(1.0, 0.93), 1.0)) = 0.93$.

Bây giờ, phép chọn trong FRDB được mở rộng từ phép chọn trong CSDL quan hệ truyền thống như sau.

Định nghĩa 10. Giả sử $R(\boldsymbol{U}, \mu)$ là một lược đồ quan hệ mờ FRDB, *r* là một quan hệ trên *R* và ϕ là một điều kiện chọn theo biến bộ *x*. *Phép chọn* trên *r* theo ϕ, được ký hiệu $\sigma_\phi(r)$, là một quan hệ mờ *r'* trên *R*, bao gồm tất cả các bộ *t* được định nghĩa bởi:

$$r' = \{t \in r \mid int_{\boldsymbol{R},r,t}(\phi) > 0 \wedge \mu_{r'}(t) = int_{\boldsymbol{R},r,t}(\phi)\}$$

Một cách đơn giản hơn, $\sigma_\phi(r) = \{t \in r \mid int_{\boldsymbol{R},r,t}(\phi) > 0\}$.

Ví dụ 6. Xét quan hệ *r* = PATIENT trong cơ sở dữ liệu các bệnh nhân ở Ví dụ 3, truy vấn "Tìm tất cả bệnh nhân trẻ và có tiền sử bệnh viêm gan" có thể được thực hiện bởi phép chọn *r'* = σ_ϕ(PATIENT) với ϕ = "*x*.AGE → *young* ∧ *x*.MEDICAL_HISTORY= *hepatitis*". Phép chọn được thực hiện bằng cách kiểm tra sự thỏa mãn của tất cả các bộ trong PATIENT đối với điều kiện chọn ϕ. Từ Ví dụ 5 ta dễ dàng thấy chỉ có bộ t_3 thỏa mãn ϕ với giá trị hàm thành viên là 0.93. Vì vậy kết quả phép chọn là quan hệ *r'* như trong Bảng 4.

IV.2. Phép hợp, giao và trừ

Sử dụng các phép toán trên các tập hợp mờ trong Định nghĩa 2, chúng tôi mở rộng các phép toán hợp, giao và trừ các quan hệ trong CSDL truyền thống thành các phép toán hợp, giao và trừ các quan hệ trong FRDB như các định nghĩa dưới đây.

Định nghĩa 11. Giả sử *r* và *s* là hai quan hệ mờ trên cùng một lược đồ $R(\boldsymbol{U}, \mu)$. *Phép hợp* (union) của hai quan hệ *r* và *s*, kí hiệu là $r \cup s$, là một quan hệ mờ trên *R* bao gồm các bộ *t* được định nghĩa bởi

$$r \cup s = \{t \mid \mu_{r\cup s}(t) = max(\mu_r(t), \mu_s(t))\}.$$

Định nghĩa 12. Giả sử *r* và *s* là hai quan hệ mờ trên cùng một lược đồ $R(\boldsymbol{U}, \mu)$. *Phép giao* (intersection) của hai quan hệ *r* và *s*, kí hiệu là $r \cap s$, là một quan hệ mờ trên *R* bao gồm các bộ *t* được định nghĩa bởi

$$r \cap s = \{t \mid \mu_{r\cap s}(t) = min(\mu_r(t), \mu_s(t))\}.$$

Định nghĩa 13. Giả sử *r* và *s* là hai quan hệ mờ trên cùng một lược đồ $R(\boldsymbol{U}, \mu)$. *Phép trừ* (difference) của quan hệ *r* cho *s*, kí hiệu là $r - s$, là một quan hệ mờ trên *R* bao gồm các bộ *t* được định nghĩa bởi

$$r - s = \{t \mid \mu_{r\cap -s}(t) = min(\mu_r(t), 1-\mu_s(t))\}.$$

IV.3. Tính chất của các phép toán đại số

Như đã thấy ở các phần trên, mô hình FRDB là được mở rộng từ mô hình CSDL quan hệ truyền thống cả về biểu diễn dữ liệu và các phép toán đại số quan hệ. Hệ quả logic là các tính chất của các phép toán đại số trong FRDB cũng được mở rộng từ các tính chất của các phép toán đại số quan hệ truyền thống.

Sau đây là các định lý về các tính chất của các phép toán đại số trên FRDB được chúng tôi mở rộng từ các tính chất của các phép toán đại số quan hệ truyền thống.

Định lý 1. Giả sử r là một quan hệ mờ trên lược đồ $R(\boldsymbol{U}, \mu)$ trong FRDB. Gọi ϕ_1 và ϕ_2 là hai điều kiện chọn. Khi đó

$$\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r) \qquad (1)$$

Với giả thiết trong $\sigma_{\phi_1 \wedge \phi_2}(r)$ các điều kiện chọn ϕ_1 và ϕ_2 là có cùng một biến bộ.

Chứng minh Đặt $s = \sigma_{\phi_2}(r)$, ta có

$\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \{t \in s \mid int_{\boldsymbol{R},s,t}(\phi_1) > 0\}$ (Định nghĩa 10)

$= \{t \in r \mid int_{\boldsymbol{R},r,t}(\phi_2) > 0 \wedge int_{\boldsymbol{R},s,t}(\phi_1) > 0\}$

$= \{t \in r \mid int_{\boldsymbol{R},r,t}(\phi_2) > 0 \wedge int_{\boldsymbol{R},r,t}(\phi_1) > 0)\}$ (do $s \subseteq r$)

$= \{t \in r \mid min(int_{\boldsymbol{R},r,t}(\phi_2), int_{\boldsymbol{R},r,t}(\phi_1)) > 0)\}$

(Định nghĩa 9)

$= \{t \in r \mid int_{\boldsymbol{R},r,t}(\phi_2 \wedge \phi_1) > 0)\} = \sigma_{\phi_1 \wedge \phi_2}(r)$.

Từ đó hệ thức $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$ được chứng minh. Hệ thức $\sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_2 \wedge \phi_1}(r)$ được chứng minh tương tự. Vì $\phi_1 \wedge \phi_2 \Leftrightarrow \phi_2 \wedge \phi_1$ (phép hội trên tập các điều kiện chọn mờ cũng như trên mệnh đề có tính giao hoán), nên $\sigma_{\phi_1 \wedge \phi_2}(r) = \sigma_{\phi_2 \wedge \phi_1}(r)$. Từ đó suy ra hệ thức $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r))$ và do đó $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$.

Định lý 2. Nếu r_1, r_2 và r_3 là các quan hệ mờ trên cùng một lược đồ $R(\boldsymbol{U}, \mu)$ thì

$$r_1 \cap r_2 = r_2 \cap r_1 \qquad (2)$$

$$(r_1 \cap r_2) \cap r_3 = r_1 \cap (r_2 \cap r_3) \qquad (3)$$

$$r_1 \cup r_2 = r_2 \cup r_1 \qquad (4)$$

$$(r_1 \cup r_2) \cup r_3 = r_1 \cup (r_2 \cup r_3) \qquad (5)$$

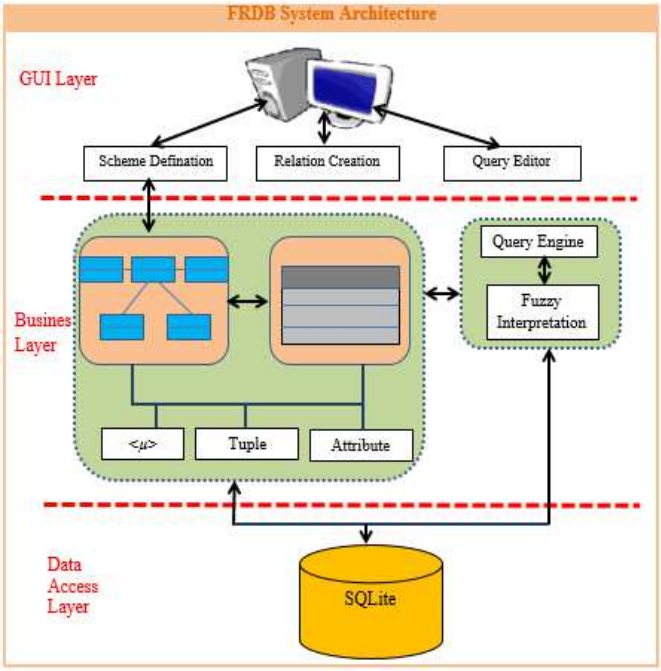
Chứng minh Do các phép toán giao và hợp các tập hợp, phép lấy *min* và *max* có tính giao hoán và kết hợp nên từ các Định nghĩa 11 và 12 ta suy các hệ thức (2), (3), (4) và (5).

V. HỆ QUẢN TRỊ FRDB

Dựa trên mô hình được đề nghị, sử dụng hệ quản trị CSDL mã nguồn mở SQLite [12] chúng tôi đã xây dựng một hệ quản trị khởi đầu cho FRDB, được gọi là FRDBS (Fuzzy Relational Database System). FRDBS cho phép thực hiện các thao tác cập nhật dữ liệu, truy vấn thông thường và truy vấn mềm với ngôn ngữ tựa SQL.

V.1. Kiến trúc của PRDBS

Kiến trúc của FRDBS gồm ba tầng như Hình 2.

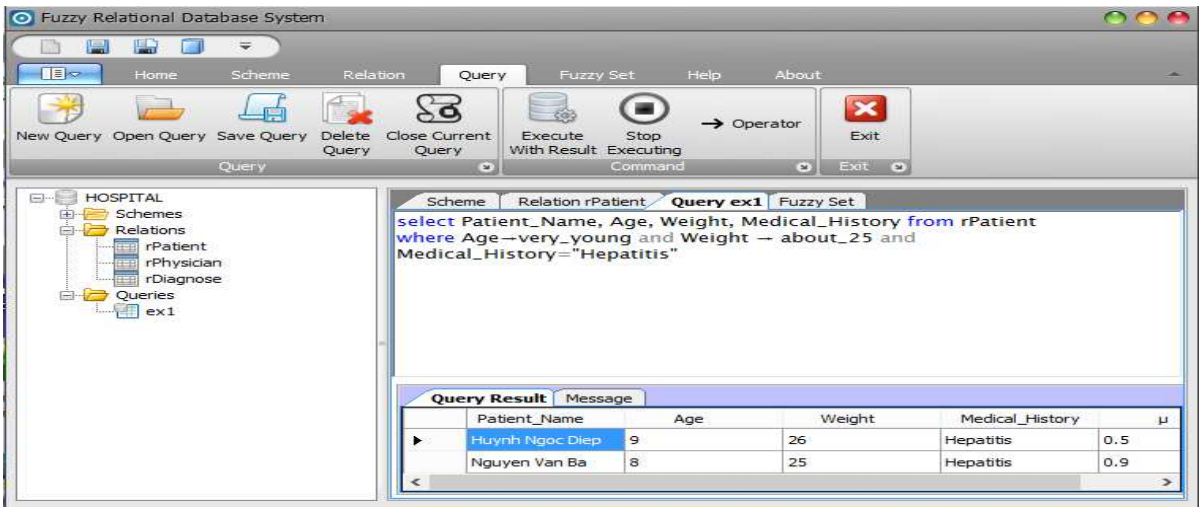


Hình 2. Kiến trúc của FRDBS

Tầng giao diện (GUI Layer): Cung cấp một giao diện đồ họa cho người dùng, cho phép người sử dụng tạo ra tập các lược đồ cùng với các thuộc tính tương ứng, từ các lược đồ đó cho phép tạo ra các quan hệ và thực hiện các truy vấn trên một quan hệ FRDB.

Tầng xử lý (Business Layer): gồm hai khối chính, một khối chịu trách nhiệm về việc mô tả, biểu diễn và lưu trữ dữ liệu của mô hình FRDB, và một khối chịu trách nhiệm thực hiện phép truy vấn. Cụ thể, khối thứ nhất sẽ đóng vai trò tạo ra cấu trúc dữ liệu của lược đồ, của quan hệ, của các kiểu giá trị thuộc tính. Khối thứ hai sẽ đóng vai trò như một "trình biên dịch", thực hiện biên dịch và thực thi câu truy vấn, cũng như thực hiện các phép diễn dịch điều kiện chọn mờ phục vụ cho việc truy vấn.

Tầng dữ liệu (Data Access Layer): là tầng giao tiếp với SQLite, cung cấp các chức năng để truy xuất, cập nhật dữ liệu cơ sở thường trú được lưu trên SQLite.



Hình 3. Một truy vấn mềm trong hệ quản trị FRDBS

V.2. Truy vấn trong PRDBS

Cú pháp truy vấn chọn trong FRDBS là tương tự cú pháp truy vấn chọn trong các hệ quản trị CSDL quan hệ truyền thống và có dạng:

select [list of attributes]
from <relation>
where [fuzzy selection condition]

Trong đó, "fuzzy selection condition" là điều kiện chọn mờ của phép chọn trên quan hệ "relation" đã được định nghĩa trong Phần IV.1. Với cú pháp này, có thể biểu diễn truy vấn mềm "tìm tất cả bệnh nhân *rất trẻ*, trọng lượng *khoảng 25* kg và có tiền sử bệnh viêm gan" trong FRDBS bởi

select Patient_Name, Age, Weight, Medical_History
from rPatient
where Age→very_young and Weight→about_25 and Medical_History="Hepatitis"

Hệ thống FRDBS thực thi truy vấn và trả về kết quả như trong Hình 3.

Chúng tôi lưu ý rằng, quan hệ mờ trên đó truy vấn thực thi là "rPatient". Đây là một tập các bộ cùng với mức độ thành viên của chúng được tạo ra từ lược đồ PATIENT. Phép toán logic "and" trong truy vấn là hiện thực của phép toán logic "∧" trong FRDB (Phần IV.1). Các tập mờ "very_young" và "about_25" đã được người dùng định nghĩa, lưu trữ trong một thư viện và được FRDBS quản lý.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã giới thiệu một mô hình cơ sở dữ liệu quan hệ mờ, được gọi là FRDB, cùng với các phép toán đại số cơ bản như chọn, hợp, giao và trừ để cho phép thao tác và truy vấn thông tin không rõ ràng, không chính xác. Mỗi quan hệ FRDB là một tập mờ với mức độ thành viên bộ trong khoảng [0, 1], các truy vấn mềm có thể được thực hiện bằng cách sử dụng các điều kiện chọn kết hợp với các giá trị tập mờ. Một số tính chất của các phép toán đại số trên FRDB cũng được đề nghị và chứng minh. Một hệ quản trị khởi đầu của FRDB được xây dựng như một minh họa cho triển vọng ứng dụng của FRDB.

Trong các bước tiếp theo, chúng tôi sẽ xây dựng các phép toán đại số khác như phép *chiếu* (projection), phép tích Descartes và phép *kết* (join) các quan hệ mờ trong FRDB và hoàn thiện hệ quản trị FRDBS để có thể mở rộng hơn nữa khả năng ứng dụng của mô hình FRDB.

TÀI LIỆU THAM KHẢO

[1] D.DUBOIS, H.PRADE, *Using fuzzy sets in flexible querying: why and how*?, Proceedings of the workshop

on flexible query-answering systems (FQAS'1996), Denmark, 1996, 89-103.

[2] E.F.CODD, *A Relational model of data for large shared data banks*, Communications of the ACM, Vol.13, No.6, 1970, 377-387.

[3] J.C.CUBERO, J.M.MEDINA, O.PONS, M.A.VILA, *Data summarization in relational databases through fuzzy dependencies*. International Journal of Information Sciences, Vol.121, 1999, 22-43.

[4] S.CHAKRABORTY, *Codd's relational data model and fuzzy logic: a practical approach to find the computer solution*. International Journal of Advanced Technology & Engineering Research (IJATER), Vol.2, No.4, 2012, 21-27.

[5] C.J.DATE, *An introduction to database systems*, Addision–Wesley, 8th Edition, 2008.

[6] G.J.KLIR, B.YUAN, *Fuzzy sets and fuzzy logic - Theory and applications*, Prentice Hall PTR, 1994.

[7] X.MENG, Z.M.MA, X.ZHU, *A Knowledge-based fuzzy query and results ranking approach for relational databases*. Journal of Computational Information Systems, Vol.6, 2010, 2037-2044.

[8] J.MISHRA, S.GHOSH, *A new functional dependency in a vague relational database Model*. International Journal of Computer Applications, Vol.39, No.8, 2012, 29-36.

[9] NGUYEN CAT HO, *A model of relational with linguistic data of hedge algebras-based semantics*, Proceedings of the 3rd National Symposium on Research, Development and Application of Information and Communication Technology (ICTrda'06) Hanoi-Vietnam, 2006, 145-156.

[10] F.E.PETRY, *Fuzzy databases: Principles and applications*, Kluwer Academic Publishers, 1996.

[11] L.YAN, Z.M.MA, *A Fuzzy probabilistic relational database model and algebra*, International Journal of Fuzzy Systems, Vol.15, No.1, 2013, 244-253.

[12] M. OWENS, *The definitive guide to SQLite*, Apress, 2006.

Nhận bài ngày: 7/10/2014

SƠ LƯỢC VỀ TÁC GIẢ

NGUYỄN HÒA

Sinh ngày 13/4/1962 tại Nghệ An.

Tốt nghiệp Đại học Sư phạm Vinh, chuyên ngành Toán học năm 1982. Nhận bằng Thạc sỹ chuyên ngành Khoa học Máy tính năm 2003 tại Trường Đại học Bách Khoa – ĐH Quốc gia TP.HCM. Nhận bằng Tiến sĩ chuyên ngành Khoa học Máy tính năm 2008 tại Trường Đại học Bách Khoa - ĐHQG TP.HCM.

Hiện công tác tại Đại học Sài gòn.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu mờ, cơ sở dữ liệu xác suất, biểu diễn và suy luận tri thức không chắc chắn, không chính xác.

Điện thoại: 0918108944, 08.38382664

Email: nguyenhoa@sgu.edu.vn