Tạp chí Tin học và Điều khiển học, T.23, S.2 (2007), 164–178

PHỤ THUỘC DỮ LIỆU TRONG CƠ SỞ DỮ LIỆU QUAN HỆ VỚI THÔNG TIN NGÔN NGỮ

NGUYỄN VĂN LONG

Khoa Công nghệ Thông tin, Đại học Giao thông Vận tải Hà Nội

Abstract. Relational databases with linguistic data based on hedge algebras - based semantics were introduced and investigated in [3], in which the evaluation of queries containing linguistic data was transformed into that of traditional queries. On this new viewpoint, in the present paper a notion of "fuzzy" functional dependencies in these databases will be defined reasonably. These new dependencies will be examined in the context of traditional functional dependencies, which play as syntaxtical constraints of the databases under consideration. Relationship between these two kinds of such dependencies will also be considered.

Tóm tắt. CSDL ngôn ngữ với ngữ nghĩa dựa trên cách tiếp cận đại số gia tử đã được nghiên cứu trong [3], trong đó việc lượng giá các truy vấn liên quan đến thông tin ngôn ngữ được đưa về việc thao tác lượng giá kinh điển. Trên cơ sở đó, phụ thuộc hàm mờ trong CSDL ngôn ngữ sẽ được định nghĩa và nghiên cứu trong ngữ cảnh với phụ thuộc hàm kinh điển và có ràng buộc CSDL ở mức cú pháp. Mối quan hệ giữa hai loại phụ thuộc này cũng được xem xét.

1. MỞ ĐẦU

Cơ sở dữ liệu (CSDL) quan hệ mờ đã được nghiên cứu phát triển từ cuối những năm 70 thể kỷ XX và từ đó đã được ứng dụng ([1, 2, 17, 21, 22]) để giải quyết các bài toán thực tiễn trong môi trường thông tin mờ, không chắc chắn. Những ứng dụng hệ thống CSDL trong thực tiễn kinh tế, xã hội thường gặp những thông tin không chắc chắn như vậy. Vì vậy, việc nghiên cứu các CSDL với thông tin mờ, không chắc chắn, không chính xác sẽ có những ứng dụng thiết thực.

Sự hiện diện các thông tin mờ, không chắc chắn trong CSDL, tất nhiên sẽ làm thay đổi căn bản việc thao tác dữ liệu cả trong phạm vi cú pháp (thao tác trên ký hiệu) và trong phạm vi ngữ nghĩa. Tuy nhiên, theo sự hiểu biết của các tác giả bài báo này, không có nhiều các công trình nghiên cứu đề cập đến những sự khác biệt sâu sắc trong phạm vi cú pháp của CSDL mờ so với CSDL kinh điển. Phần lớn các nghiên cứu các phụ thuộc dữ liệu (PTDL) trong CSDL mờ đều là sự mở rộng của các PTDL kinh điển, nghĩa là các PTDL đó vẫn đúng khi các dữ liệu trong CSDL đều là thực. Trong những trường hợp như vậy, đối với CSDL mờ, chúng ta đã không mở rộng được cú pháp của lớp các PTDL, và do đó không ảnh hưởng đến việc thiết kế CSDL. Khi đó, ta chỉ mở rộng được ngữ nghĩa hay các quan hệ ngữ nghĩa của các dữ liệu để cho phép khai thác dữ liệu trong CSDL mờ.

Nhờ những ưu việt của cấu trúc đại số gia tử (ĐSGT) ([4–16, 18, 19]), trong [3] đã đưa ra và nghiên cứu CSDL mờ dựa trên cách tiếp cận của đại số gia tử, trong đó ngữ nghĩa ngôn

ngữ được lượng hóa bằng các ánh xạ định lượng của ĐSGT. Theo cách tiếp cận này, giá trị ngôn ngữ là dữ liệu, không phải là nhãn của các tập mờ biểu diễn ngữ nghĩa của giá trị ngôn ngữ và ưu điểm cơ bản của nó là cho phép tìm kiếm, đánh giá ngữ nghĩa của thông tin không chắc chắn chỉ bằng các thao tác dữ liệu kinh điển thường dùng và do đó bảo đảm tính thuần nhất của kiểu dữ liệu trong xử lý ngữ nghĩa của chúng. Điều này khác với CSDL mờ là vừa phải xử lý ngữ nghĩa kinh điển, vừa phải xử lý ngữ nghĩa được biểu diễn dưới dạng các tập mờ hay hàm thuộc của chúng. Theo cách tiếp cận của ĐSGT, ngữ nghĩa ngôn ngữ có thể biểu thị bằng một lân cận các khoảng được xác định bởi độ đo tính mờ của các giá trị ngôn ngữ của một thuộc tính với vai trò là biến ngôn ngữ. Ví dụ, ngữ nghĩa của giá trị ngôn ngữ (GTNNg) rất lớn của thuộc tính "Số bài trên tạp chí nước ngoài" sẽ được biểu thị bằng những khoảng lân cận của giá trị đại diện của GTNNg rất lớn thông qua ánh xạ định lượng của ĐSGT của thuộc tính "Số bài trên tạp chí nước ngoài". Theo nghĩa đó, trong [3] đã sử dụng thuật ngữ CSDL ngôn ngữ thay cho thuật ngữ CSDL mờ.

Bài báo này sẽ nghiên cứu các PTDL mờ trong CSDL ngôn ngữ trên cả hai khía cạnh cú pháp (syntax) và ngữ nghĩa (semantics). Ta sẽ thấy trong CSDL ngôn ngữ vừa tồn tại các PTDL mở rộng kinh điển, vừa tồn tại những PTDL mờ, tức là không có sự PTDL kinh điển được biểu thị bằng các PTDL mờ này.

Tiếp theo, những khái niệm cơ bản về ĐSGT và CSDL ngôn ngữ sẽ được trình bày ngắn gọn trong Mục 2, đặc biệt khái niệm về độ tương tự giữa các dữ liệu của thuộc tính ngôn ngữ sẽ được đề cập lại. Trong Mục 3, khái niệm phụ thuộc hàm tương tự sẽ được định nghĩa và nghiên cứu. Nó sẽ là một sự mở rộng rất gần gũi với phụ thuộc hàm kinh điển và do vậy chúng có thể được nghiên cứu trong mối liên hệ chặt chẽ với nhau. Hệ tiên đề Armstrong và tính đầy đủ của nó vẫn còn đúng đối với lớp phụ thuộc mới và do đó vai trò khác biệt của hai loại phụ thuộc dữ liệu này trong cùng một CSDL ngôn ngữ cũng được xem xét. Một số kết luận và các vấn đề ngỏ được trình bày trong phần kết luận, Mục 4.

2. NHỮNG KHÁI NIỆM CƠ BẢN VỀ ĐSGT VÀ CSDL VỚI THÔNG TIN NGÔN NGỮ

2.1. Về đại số gia tử (ĐSGT)

Để dễ theo dõi phương pháp xử lý ngữ nghĩa ngôn ngữ theo cách tiếp cận ĐSGT, ta tóm tắt lại một số khái niệm về ánh xạ định lượng và cách thức xác định các hệ lân cận ngữ nghĩa định lượng. Trong CSDL ngôn ngữ, các thuộc tính ngôn ngữ, ngoài các giá trị kinh điển chúng có thể có các ký hiệu giá trị ngôn ngữ. Vì vậy, mỗi thuộc tính ngôn ngữ sẽ được gắn kết với một ĐSGT.

Cho một ĐSGT tuyến tính đầy đủ $\mathcal{AX} = (\boldsymbol{X}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{\sigma}, \boldsymbol{\Phi}, \leqslant)$, trong đó $Dom(\mathcal{X}) = \boldsymbol{X}$ là miền các giá trị ngôn ngữ của thuộc tính ngôn ngữ $\mathcal{X}$ được sinh tự do từ tập các phần thử sinh $G = \{\mathbf{1}, c^+, \boldsymbol{W}, c^-, \mathbf{0}\}$ bằng việc tác động tự do các phép toán một ngôi (các gia tử) trong tập $\boldsymbol{H}$; $\boldsymbol{\sigma}$ và $\boldsymbol{\phi}$ là hai phép tính với ngữ nghĩa là cận trên đúng và cận dưới đúng của tập $\boldsymbol{H}(x)$, tức là $\sigma x = supremum\boldsymbol{H}(x)$ and $\phi x = infimum\boldsymbol{H}(x)$, trong đó $\boldsymbol{H}(x)$ là tập các phần từ sinh ra từ x, còn quan hệ $\leqslant$ là quan hệ sắp thứ tự tuyến tính trên $\boldsymbol{X}$ cảm sinh từ ngữ nghĩa của ngôn ngữ. Ví dụ, nếu ta có thuộc tính $Num\mathcal{IP}$ (Number of International Papers) là "Số bài báo đăng trên tạp chí quốc tế", thì $Dom(Num\mathcal{IP}) =$

$\{large, small, verylarge, morelarge, possiblylarge, verysmall, possiblysmall, lesssmall, ...\}$, $\boldsymbol{G} = \{\mathbf{1}, large, \boldsymbol{W}, small, \mathbf{0}\}$, $\boldsymbol{H} = \{very, more, possibly, little\}$ và $\leqslant$ một quan hệ thứ tự cảm sinh từ ngữ nghĩa của các từ trong $Dom(Num\mathcal{IP})$, chẳng hạn ta có $verylarge > large, morelarge > large, possiblylarge < large, littlelarge < large, ...$

Dựa trên cấu trúc của ĐSGT, trong đó quan hệ giữa các phần tử là quan hệ thứ tự ngữ nghĩa, mô hình toán học của tính mờ và độ đo tính mờ của các khái niệm mờ đã được định nghĩa trong [6, 7].

Giả sử các gia tử trong tập $\boldsymbol{H} = \boldsymbol{H}^- \cup H^+$, được liệt kê như sau:

$\boldsymbol{H}^+ = \{h_1, ..., h_p\}$ và $\boldsymbol{H}^- = \{h_{-1}, ..., h_{-q}\}$, với $h_1 < ... < h_p$ và $h_{-1} < ... < h_{-q}$, trong đó $p, q > 1$.

Cho $fm : X \to [0, 1]$ là độ đo tính mờ của ĐSGT $\mathcal{AX}$, ta có mệnh đề sau.

Mệnh đề 2.1. ([6, 7]) *Độ đo tính mờ fm và độ đo tính mờ của gia tử $\mu(h), \forall h \in \boldsymbol{H}$, có các tính chất sau:*

1) $fm(hx) = \mu(h) fm(x),\ \forall x \in \boldsymbol{X}$

2) $fm(c^-) + fm(c^+) = 1$

3) $\sum\limits_{-q\ \ i\ \ p,\ i\neq 0} fm(h_i c) = fm(c)$, *trong đó* $c \in \{c^-, c^+\}$

4) $\sum\limits_{-q\ \ i\ \ p,\ i\neq 0} fm(h_i x) = fm(x)$, $x \in \boldsymbol{X}$

5) $\sum\{\mu(h_i) : -q \leqslant i \leqslant -1\} = \alpha$ *và,* $\sum\{\mu(h_i) : 1 \leqslant i \leqslant p\} = \beta$, *trong đó* $\alpha + \beta > 0$ *và* $\alpha + \beta = 1$.

Ở đây mặc dù 3) là trường hợp riêng của 4), nhưng vẫn được viết ra để dễ hình dung việc hình thành các khoảng tính mờ của các khái niệm mờ.

Khoảng mờ của khái niệm mờ. Giả sử thuộc tính (hay biến ngôn ngữ) $\mathcal{X}$ có miền tham chiếu thực là khoảng $[a, b]$. Để chuẩn hóa, nhờ một phép biến đổi tuyến tính, ta giả thiết mọi miền như vậy đều là khoảng $[0, 1]$. Vì độ đo tính mờ fm là một ánh xạ $\boldsymbol{X} \to [0, 1]$, nên nó gợi ý đến việc biểu diễn các giá trị $fm(x),\ x \in \boldsymbol{X}$, bằng các khoảng con của đoạn $[0, 1]$ và được gọi là khoảng mờ của khái niệm x. Như vậy, các khoảng mờ là một biểu diễn định lượng các khái niệm mờ của một biến ngôn ngữ. Cho trước fm, các khoảng mờ của các khái niệm mờ trong $\boldsymbol{X}$ được xây dựng quy nạp theo độ dài của $x \in \boldsymbol{X}$ như sau:

- Khoảng mờ của hai khái niệm nguyên thủy c^- và c^+: Rõ ràng là từ tính chất 2) ta có thể xây dựng hai khoảng mờ $\Im(c^-)$ và $\Im(c^+)$ của hai khái niệm nguyên thủy c^- và c^+, với $|\Im(c^-)| = fm(c^-)$ và $|\Im(c^+)| = fm(c^+)$, trong đó $|\Im(x)|$ chỉ độ dài của khoảng $\Im(x)$, sao cho và chúng tạo thành một phân hoạch của $[0, 1]$ và $\Im(c^-)$, $\Im(c^+)$ đồng biến với c^-, c^+, tức là $c^- \leqslant c^+$ kéo theo $\Im(c^-) \leqslant \Im(c^+)$, ở đây $\Im(c^-) \leqslant \Im(c^+)$ được hiểu là với $\forall x \in \Im(c^-)$ và $\forall y \in \Im(c^+)$, ta có $x \leqslant y$.

- Khoảng mờ của x độ dài $k > 1$: Một cách quy nạp, ta giả sử rằng với $\forall x \in \boldsymbol{X}_{k-1} = \{x \in X : x \text{ có độ dài } |x| = k-1\}$, ta đã xây dựng được khoảng mờ $\Im(x)$, với $|((x)| = fm(x)$, sao cho $\{\Im(x) : x \in \boldsymbol{X}_{k-1}\}$ đồng biến với thứ tự trên tập $\boldsymbol{X}_{k-1}$ và tạo thành một phân hoạch của đoạn $[0, 1]$. Khi đó, trên mỗi khoảng mờ $\Im(x)$ của $x \in \boldsymbol{X}_{k-1}$, do tính chất 4), ta có thể xây dựng được họ các khoảng $\{\Im(h_i x) : q \leqslant i \leqslant p,\ i \neq 0,\ |\Im(h_i x)| = fm(h_i x)\}$ sao cho chúng là một phân hoạch của khoảng mờ $\Im(x)$ và đồng biến với thứ tự các phần tử $\{h_i x : q \leqslant i \leqslant p,\ i \neq 0\}$.

Có thể thấy họ $\{\Im(h_i x) : q \leqslant i \leqslant p,\ i \neq 0,\ |\Im(h_i x)| = fm(h_i x)$ và $x \in \boldsymbol{X}_{k-1}\} = \{\Im(y) :$

$y \in \boldsymbol{X}_k$ và $|\Im(y)| = fm(y)\}$ là một phân hoạch của $[0,1]$. Các khoảng này gọi là các khoảng mờ mức k.

Như vậy, có một sự liên hệ chặt chẽ giữa ngữ nghĩa ngôn ngữ của các khái niệm mờ và các khoảng mờ trong đoạn $[0,1]$ như sau: (i) Mỗi phần tử $x \in \boldsymbol{X}$ đều được gắn với một khoảng mờ $\Im(x)$ có độ dài chính bằng độ đo tính mờ của x; (ii) Nếu x' là hậu tố của x, tức là nó là xâu con bên trái cùng của xâu x hay, nói khác đi, x' sinh ra xâu x, thì $\Im(x) \subset \Im(x')$; (iii) Nếu x và x' có cùng độ dài và $x \leqslant x'$ thì $\Im(x) \leqslant \Im(x')$.

Đây là những tính chất rất quan trọng của các khoảng mờ được định nghĩa dựa trên cấu trúc ĐSGT và là cơ sở để định nghĩa hệ lân cận ngữ nghĩa của x đã được định nghĩa trong [3] và sẽ được trình bày tóm tắt dưới đây.

Định nghĩa 2.1. Ánh xạ $f : X \to [0,1]$ được gọi là ánh xạ định lượng của ĐSGT $\mathcal{AX}$ nếu nó thỏa mãn các điều kiện sau:

Q1) f là ánh xạ đơn ánh.

Q2) f bảo toàn quan hệ thứ tự ngữ nghĩa trên $\boldsymbol{X}$, nghĩa là $x < y \Rightarrow f(x) < f(y)$, và $f(\mathbf{0}) = 0,\ f(\mathbf{1}) = 1$.

Q3) f liên tục theo nghĩa với $\forall x \in \boldsymbol{X},\ f(\boldsymbol{\phi}x) = \inf f(\boldsymbol{H}(x))$ và $f(\boldsymbol{\sigma}x) = \sup f(\boldsymbol{H}(x))$.

Trong đại số gia tử, mỗi phần tử $x \in \boldsymbol{X}$ đều mang dấu âm hay dương, được gọi là PN-dấu và được định nghĩa đệ quy như sau.

Định nghĩa 2.2. (Hàm PN-dấu Sgn) $Sgn : X \to \{-1, 0, 1\}$ là hàm dấu được xác định như sau, ở đây $h, h' \in \boldsymbol{H}$, và $c \in \{c^-, c^+\}$

a) $Sgn(c^-) = -1,\ Sgn(c^+) = +1$.

b) $Sgn(h'hx) = 0$, nếu $h'hx = hx$, ngược lại ta có:

$Sgn(h'hx) = -Sgn(hx)$, nếu $h'hx \neq hx$ và h' là âm tính đối với h (hoặc c, nếu $h = I$ và $x = c$),

$Sgn(h'hx) = +Sgn(hx)$, nếu $h'hx \neq hx$ và h' dương tính đối với h (hoặc c, nếu $h = I$ và $x = c$).

Ý nghĩa của PN-dấu thể hiện trong mệnh đề dưới đây.

Mệnh đề 2.2. *Với mọi $x \in \boldsymbol{X}$, $\forall h \in H$, nếu $Sgn(hx) = +1$ thì $hx > x$, nếu $Sgn(hx) = -1$ thì $hx < x$ và nếu $Sgn(hx) = 0$ thì $hx = x$.*

Với các tính chất của tính mờ và hàm PN-dấu, ánh xạ ngữ nghĩa định lượng của ĐSGT được định nghĩa như sau.

Định nghĩa 2.3. Giả sử $\mathcal{AX} = (\boldsymbol{X}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{\sigma}, \boldsymbol{\Phi}, \leqslant)$ là một ĐSGT đầy đủ, tuyến tính và tự do, $fm(x)$ và $\mu(h)$ tương ứng là các độ đo tính mờ của ngôn ngữ và của gia tử h thỏa mãn các tính chất trong Mệnh đề 2.1. Khi đó, ta nói ν là ánh xạ cảm sinh bởi độ đo tính mờ fm của ngôn ngữ nếu nó được xác định như sau:

1) $\nu(\boldsymbol{W}) = \kappa = fm(c^-),\ \nu(c^-) = \kappa - \alpha fm(c^-) = \beta fm(c^-),\ \nu(c^+) = \kappa + \alpha fm(c^+)$.

2) $\nu(h_j x) = \nu(x) + Sgn(h_j x)\{\sum_{i=Sgn(j)}^{j} \mu(h_i) fm(x) - \omega(h_j x)\mu(h_j) fm(x)\}$ trong đó,

$\omega(h_j x) = \frac{1}{2}[1 + Sgn(h_j x) Sgn(h_p h_j x)(\beta - \alpha)] \in \{\alpha, \beta\}$, với mọi $j, -q \leqslant j \leqslant p$ và $j \neq 0$.

3) $\nu(\boldsymbol{\phi}c^-) = 0,\ \nu(\sigma(c^-) = \kappa = \nu(\boldsymbol{\phi}c^+),\ \nu(\sigma c^+) = 1$, và với mọi $j, -q \leqslant j \leqslant p$ và $j \neq 0$, chúng ta có:

$$\nu(\phi h_j x) = \nu(x) + Sgn(h_j x)\{\sum_{i=Sign(j)}^{j-1} \mu(h_i) fm(x)\} \text{ và}$$

$$\nu(\sigma h_j x) = \nu(x) + Sgn(h_j x)\{\sum_{i=Sign(j)}^{j} \mu(h_i) fm(x)\}.$$

Thức chất của ánh xạ ν là, với mọi $x = h_j u$, $\nu(x)$ chính là điểm chia trong khoảng mờ $\Im(x)$ theo tỷ lệ $\alpha : \beta$ nếu $Sgn(h_p h_j u) = +1$ và, nếu ngược lại, nó là điểm chia trong khoảng $\Im(x)$ theo tỷ lệ $\beta : \alpha$. Như vậy, chúng ta có mối quan hệ chặt chẽ giữa giá trị ánh xạ định lượng của ĐSGT và ngữ nghĩa của giá trị ngôn ngữ được biểu thị qua các khoảng mờ của chúng. Điều này đặc biệt có ý nghĩa nếu chúng ta nhớ rằng bản chất các phương pháp khử mờ (defuzzification methods) chính là thiết lập các ánh xạ định lượng. Tuy nhiên, các phương pháp như vậy không dựa trên sự liên hệ mật thiết với ngữ nghĩa ngôn ngữ (!).

Chú ý: Từ Mệnh đề 2.1 và Định nghĩa 2.3 cho thấy, với bất kỳ một ĐSGT, ta luôn xây dựng được độ đo tính mờ và hàm định lượng cảm sinh từ độ đo tính mờ bằng cách lựa chọn các các giá trị độ đo tính mờ của các phần tử nguyên thủy c^- và c^+ và của các gia tử sao cho chúng thỏa mãn tính chất 2) và 5) của Mệnh đề 2.1, còn gọi là bộ tham số định lượng. Nói khác đi, ta sẽ có một bộ tham số để điều chỉnh cho thích ứng với một ứng dụng cụ thể nào đó. Ngoài ra, từ tính chất 3), Định nghĩa 2.3 ta thấy, khoảng mờ $\Im(x) = (\nu_A(\phi x),\ \nu_A(\sigma x)]$.

2.2. Về CSDL ngôn ngữ

Trong [3], khái niệm CSDL ngôn ngữ đã được đưa ra và nghiên cứu. Ở mức cú pháp, khái niệm này không có nhiều thay đổi và vì vậy chúng ta vẫn sử dụng các ký pháp truyền thống. Sau đây sẽ giới thiệu tóm tắt các khái niệm chính.

Xét một lược đồ CSDL $DB = \{U, R_1, R_2, ..., R_m; \text{const}\}$, trong đó $U = \{A_1, A_2, ..., A_n\}$ là tập vũ trụ các thuộc tính, R_i lược đồ quan hệ, tức là một tập con của U, const là một tập các ràng buộc dữ liệu của CSDL. Mỗi thuộc tính A được gắn với một miền giá trị thuộc tính, ký hiệu là $Dom(A)$, trong đó một số thuộc tính cho phép nhận các giá trị ngôn ngữ trong lưu trữ trong CSDL hay trong các câu hỏi truy vấn và được gọi là thuộc tính ngôn ngữ. Những thuộc tính còn lại được gọi là thuộc tính thực hay kinh điển. Thuộc tính thực A được gắn với một miền giá trị kinh điển, ký hiệu là D_A. Thuộc tính ngôn ngữ A sẽ được gắn một miền giá trị kinh điển D_A và một miền giá trị ngôn ngữ LD_A hay là tập các phần tử của một ĐSGT. Để bảo đảm tính nhất quán trong xử lý ngữ nghĩa dữ liệu trên cơ sở thống nhất kiểu dữ liệu của thuộc tính ngôn ngữ, mỗi thuộc tính ngôn ngữ sẽ được gắn với một ánh xạ định lượng $\nu_A : LD_A \to D_A$ được xác định bởi một bộ tham số định lượng của A. Như vậy, mỗi giá trị ngôn ngữ x của A sẽ được gán một nhãn giá trị thực $\nu_A(x) \in D_A$ được xem như giá trị đại diện của x. Một CSDL như vậy được gọi là CSDL ngôn ngữ.

Việc đánh giá độ tương tự (similarity degree) giữa các dữ liệu của một thuộc tính A được dựa trên khái niệm lân cận mức k của một giá trị ngôn ngữ, với k là số nguyên dương.

Độ tương tự mức k ([3]) (Similarity degree): Giả sử, ta có thể lấy các khoảng mờ của các phần tử độ dài k làm độ tương tự giữa các phần tử, nghĩa là các phần tử mà các giá trị đại diện của chúng thuộc cùng một khoảng mờ mức k là tương tự mức k. Tuy nhiên, theo cách xây dựng các khoảng mờ mức k, giá trị đại diện của các phần từ x có độ dài nhỏ hơn k luôn luôn là đầu mút của các khoảng mờ mức k. Một cách hợp lý, khi định nghĩa lân cận mức k chúng ta mong muốn các giá trị đại diện như vậy phải là điểm trong (theo nghĩa tôpô) của

lân cận mức k. Vì vậy ta định nghĩa độ tương tự mức k như sau.

Giả thiết rằng mỗi tập $\boldsymbol{H}^-$ và $\boldsymbol{H}^+$ chứa ít nhất 2 gia tử. Xét $\boldsymbol{X}_k$ là tập tất cả các phần tử độ dài k. Dựa trên các khoảng mờ mức k và các khoảng mờ mức k+1 chúng ta mô tả không hình thức ([3]) việc xây dựng một phân hoạch của miền $[0,1]$ như sau (nhớ rằng bằng một phép biến đổi tuyến tính, đoạn $[0,1]$ trở về miền thực D_A của thuộc tính A).

Với $k=1$, các khoảng mờ mức 1 gồm $\Im(c^-)$ và $\Im(c^+)$. Các khoảng mờ mức 2 trên khoảng $\Im(c^-)$ là $\Im(h_pc^-) \leqslant \Im(h_{p-1}c^-) \leqslant \cdots \leqslant \Im(h_2c^-) \leqslant \Im(h_1c^-) \leqslant \nu_A(c^-) \leqslant \Im(h_{-1}c^-) \leqslant \Im(h_{-2}c^-) \leqslant \cdots \leqslant \Im(h_{-q+1}c^-) \leqslant \Im(h_{-q}c^-)$. Khi đó, ta xây dựng phân hoạch về độ tương tự mức 1 gồm các lớp tương đương sau:

$$S(\mathbf{0}) = \Im(h_pc^-),\ S(c^-) = \Im(c^-)\backslash[\Im(h_{-q}c^-) \cup \Im(h_pc^-)],\ S(\boldsymbol{W}) = \Im(h_{-q}c^-) \cup \Im(h_{-q}c^+)$$

và, một cách tương tự, $S(c^+) = \Im(c^+)\backslash[\Im(h_{-q}c^+) \cup \Im(h_pc^+)]$ và $S(\mathbf{1}) = \Im(h_pc^+)$.

Ta thấy, trừ hai điểm đầu mút $\nu_A(\mathbf{0}) = 0$ và $\nu_A(\mathbf{1}) = 1$, các giá trị đại diện $\nu_A(c^-),\ \nu_A(\boldsymbol{W})$ và $\nu_A(c^+)$ đều là điểm trong tương ứng của các lớp tương tự mức 1 $S(c^-),\ S(\boldsymbol{W})$ và $S(c^+)$.

Tương tự, với $k=2$, ta có thể xây dựng phân hoạch các lớp tương tự mức 2. Chẳng hạn, trên một khoảng mờ mức 2, $\Im(h_ic^+) = (\nu_A(\boldsymbol{\phi}h_ic^+),\ \nu_A(\sigma h_ic^+)]$ với hai khoảng mờ kề là $\Im(h_{i-1}c^+)$ và $\Im(h_{i+1}c^+)$, ta sẽ có các lớp tương đương dạng $S(h_ic^+) = \Im(h_ic^+)\backslash[\Im(h_ph_ic^+) \cup \Im(h_{-q}h_ic^+)],\ S(\boldsymbol{\phi}h_ic^+) = \Im(h_{-q}h_{i-1}c^+) \cup \Im(h_{-q}h_ic^+)$ và $S(\boldsymbol{\phi}h_ic^+) = \Im(h_ph_ic^+) \cup \Im(h_ph_ic^+)$, với i sao cho $-q \leqslant i \leqslant p$ và $i \neq 0$.

Bằng cách tương tự, ta có thể xây dựng các phân hoạch các lớp tương tự mức k bất kỳ. Tuy nhiên, trong thức tế thì $k < 4$, tức có tối đa 3 gia tử tác động liên tiếp lên phần tử nguyên thủy c^- và c^+. Các giá trị kinh điển và các giá trị ngôn ngữ được gọi là có độ tương tự mức k nếu các giá trị đại diện của chúng (ở đây đại diện của giá trị thực là chính nó) cùng nằm trong một lớp tương tự mức k.

Lân cận mức k của khái niệm mờ: Giả sử phân hoạch các lớp tương tự mức k là các khoảng $S(x_1), S(x_2), ..., S(x_m)$. Khi đó, mỗi giá trị ngôn ngữ $\boldsymbol{u}$ chỉ và chỉ thuộc về một lớp tương tự, chẳng hạn đó là $S(x_i)$ và được gọi là lân cận mức k của $\boldsymbol{u}$, ký hiệu là $\Omega_k(u)$.

Dựa trên khái niệm độ tương tự như vậy, các quan hệ đối sánh được định nghĩa như sau.

Định nghĩa 2.4. ([3]) Giả sử t và s là hai bộ dữ liệu trên tập vũ trụ U các thuộc tính. Ta nói $t[A_i] =_{k(A_i)} s[A_i]$ và gọi chúng là bằng nhau mức k, nếu một trong các điều kiện sau xảy ra:

(i) Nếu $t[A_i],\ s[A_i] \in D_{A_i}$ thì $t[A_i] = s[A_i]$.

(ii) Nếu một trong hai giá trị $t[A_i],\ s[A_i]$ là khái niệm mờ, chẳng hạn đó là $t[A_i]$, thì ta phải có $s[A_i] \in \Omega_k(t[A_i])$.

(iii) Nếu cả hai giá trị $t[A_i],\ s[A_i]$ đều là ngôn ngữ, thì $\Omega_k(t[A_i]) = \Omega_k(s[A_i])$.

Như thông lệ, nếu điều kiện $t[A_i] =_{k(A_i)} s[A_i]$ không xảy ra, ta có biểu thức:

$$t[A_i] \neq_{k(A_i)} s[A_i].$$

Do quan hệ tương tự mức k được xây dựng bằng một phân hoạch của đoạn $[0,1]$, nên có thể thấy quan hệ $=_{k(A_i)}$ là tương đương trên $[0,1]$. Ngoài ra, ta cần nhấn mạnh rằng đẳng thức $t[A_i] =_{k(A_i)} s[A_i]$ có nghĩa $L_k \leqslant t[A_i],\ s[A_i] \leqslant R_k$, trong đó L_k và R_k là hai điểm mút của khoảng $\Omega_k(t[A_i])$ hay $\Omega_k(s[A_i])$, nghĩa là, việc kiểm chứng $t[A_i] =_{k(A_i)} s[A_i]$ được đưa về việc kiểm chứng các quan hệ đối sánh kinh điển. Hơn nữa, tính mềm dẻo trong thích nghi

với các ứng dụng cụ thể có thể đạt được bằng việc điều chỉnh các tham số của ánh xạ định lượng ν_{A_i}. Đây chính là ưu điểm nổi bật của cách tiếp cận đại số đến thông tin ngôn ngữ.

Dựa trên quan hệ tương đương này ta có thể dễ dàng định nghĩa các quan hệ đối sánh khác. Trước hết, để đơn giản ta quy ước là ký pháp $\Omega_k(t[A_i])$ có nghĩa cả khi $t[A_i] \in D_{A_i}$. Khi đó $\Omega_k(t[A_i])$ được hiểu là tập bao gồm chỉ đúng một giá trị thực $t[A_i]$. Với quy ước đó, với mọi cặp lân cận mức k, $\Omega_k(x)$ and $\Omega_k(y)$, ta sẽ viết $\Omega_k(x) < \Omega_k(y)$ khi $u < v$, với mọi $u \in \Omega_k(x)$ và mọi $\nu \in \Omega_k(y)$.

Định nghĩa 2.5. Giả sử t và s là hai bộ dữ liệu trên tập vũ trụ U các thuộc tính. Khi đó,

(i) $t[A_i] \leqslant_{\nu,k} s[A_i]$, nếu $t[A_i] =_{\nu,k} s[A_i]$ hoặc $\Omega_k(t[A_i]) < \Omega_k(s[A_i])$.

(ii) $t[A_i] <_{\nu,k} s[A_i]$, nếu $\Omega_k(t[A_i]) < \Omega_k(s[A_i])$.

(ii) $t[A_i] >_{\nu,k} s[A_i]$, nếu $\Omega_k(t[A_i]) > \Omega_k(s[A_i])$.

Các quan hệ xác định trong Định nghĩa 2.4 và 2.5 được gọi là các quan hệ đối sánh mờ trong CSDL ngôn ngữ.

3. PHỤ THUỘC HÀM DỰA TRÊN ĐỘ TƯƠNG TỰ TRONG CSDL NGÔN NGỮ

Xét CSDL ngôn ngữ $DB = \{U; \text{const}\}$, trong đó $U = \{A_1, A_2, ..., A_n\}$. Như trên đã đề cập, const là tập các ràng buộc dữ liệu. Việc phân tích một quan hệ vũ trụ trên U thành các quan hệ $R_1, R_2, ..., R_m$ nhỏ hơn được dựa trên lý thuyết thiết kế CSDL trong đó phụ thuộc hàm, phụ thuộc đa trị, phụ thuộc kết nối đóng vai trò quan trọng. Một khi ngữ nghĩa của CSDL được mở rộng, như cho phép lưu giữ trong CSDL các thông tin không chắc chắn hay cho phép các câu hỏi truy vấn chứa các thông tin như vậy, thì ngữ nghĩa của các phụ thuộc dữ liệu cũng thay đổi, nghĩa là phải mở rộng định nghĩa phụ thuộc dữ liệu.

Trong thực tế, ta thường thu được các tri thức dạng không chắc chắn như: *Nếu một tập thể khoa học có nhiều công trình nghiên cứu khoa học công bố trên các tạp chí có giá trị, nhiều cán bộ được đào tạo từ nghiên cứu viên thành Tiến sỹ thì NĂNG LỰC TỔ CHỨC hoạt động khoa học công nghệ được đánh giá xuất sắc.* Ở đây ta không nhìn nhận mối quan hệ trên như là một luật của một cơ sở tri thức nào đó mà xem như là mối quan hệ giữa các thuộc tính trong CSDL với thuộc tính SỐ CÔNG TRÌNH, SỐ CÁN BỘ TRƯỞNG THÀNH và NĂNG LỰC TỔ CHỨC. Tuy nhiên mối quan hệ này không chính xác như mối quan hệ phụ thuộc kinh điển, kể cả khi hai thuộc tính đầu trong CSDL chỉ lưu các giá trị kinh điển. Vì vậy, mục tiêu của bài báo này là nghiên cứu các phụ thuộc dữ liệu với thông tin không chắc chắn.

Dựa vào cách tiếp cận đại số đối với ngữ nghĩa của giá trị ngôn ngữ của các thuộc tính ngôn ngữ, chúng tôi đưa một khái niệm phụ thuộc hàm mờ, gọi là *phụ thuộc hàm tương tự*.

Trong thực tế có nhiều sự phụ thuộc giữa các tiêu chí trên cơ sở thông tin không chính xác hay thông tin ngôn ngữ. Ví dụ trong đánh giá năng lực tổ chức hoạt động nghiên cứu khoa học có tồn tại sự phụ thuộc giữa thuộc tính SỐ CÔNG TRÌNH CÔNG BỐ NƯỚC NGOÀI, SỐ CÔNG TRÌNH CÔNG BỐ TRONG NƯỚC, SỐ CÁN BỘ TRƯỞNG THÀNH TRONG NGHIÊN CỨU (số lượng nhận học vị ThS hay TS) và NĂNG LỰC TỔ CHỨC (đánh giá bằng thang bậc ngôn ngữ hay điểm). Ta thấy không thể có sự phụ thuộc chính xác giữa các tiêu chí này, dù rằng dữ liệu là kinh điển. Trong mối quan hệ phụ thuộc như vậy, có tồn tại

những phụ thuộc hàm "không chắc chắn" trên cơ sở các quan sát sau.

Ta xét một tình hình thực tế trong đánh giá của một giảng viên về năng lực tư duy của sinh viên về môn toán trên cơ sở các môn học GIẢI TÍCH, ĐẠI SỐ và HÌNH HỌC và NĂNG LỰC TƯ DUY cho bằng thang bậc ngôn ngữ như "suất sắc", "khá", "trung bình" và "kém". Thực tế là thuộc tính NĂNG LỰC TƯ DUY được định giá trong nhận thức của người giảng viên đó (hay của bất kỳ ai đã tham gia đào tạo) không nhất thiết là một phụ thuộc hàm kinh điển vào các thuộc tính còn lại, vì có thể có hai sinh viên có cùng các điểm đánh giá các môn học nhưng được đánh giá năng lực tư duy khác nhau. Hay đối với vấn đề hướng nghiệp cho học sinh, cũng không nhất thiết các kết quả học tập như nhau, hoàn cảnh như nhau nhưng lời khuyên hướng nghiệp như nhau vì nó còn phụ thuộc vào những đánh giá trực quan khác nhưng chưa hoặc khó có tiêu chí đánh giá. Ngoài ra, bản chất những mối quan hệ phụ thuộc như vậy không chắc chắn hay "không có ranh giới rõ ràng" cho việc xác định các phụ thuộc.

Trong CSDL ngôn ngữ, có thể hình thức hóa một lớp các phụ thuộc không chắc chắn kiểu như trên, được gọi là phụ thuộc tương tự vì nó dựa trên sự đối sánh tương tự đã được hình thức hóa ở trên.

Trước khi ta đưa ra định nghĩa phụ thuộc hàm tương tự, một câu hỏi cần giải đáp là với các thuộc tính ngôn ngữ khác nhau, mức k của độ tương tự trong một ứng dụng có giống nhau hay không? Để trả lời, ta quan sát và thấy, trong trường hợp cực đoan là giữa thuộc tính kinh điển và thuộc tính ngôn ngữ, các mức của độ tương tự là khác nhau, vì mức của độ tương tự của thuộc tính kinh điển có mức 0, nghĩa là chúng đồng nhất bằng nhau, còn mức của độ tương tự của thuộc tính ngôn ngữ là khác 0. Với cách nhìn đó, nếu bắt buộc trên các thuộc tính SỐ CÔNG TRÌNH CÔNG BỐ NƯỚC NGOÀI, SỐ CÔNG TRÌNH CÔNG BỐ TRONG NƯỚC, SỐ CÁN BỘ TRƯỞNG THÀNH và NĂNG LỰC TỔ CHỨC trong mối quan hệ phụ thuộc dữ liệu nêu trên đều có độ tương tự mức k như nhau là một ràng buộc không thực tế.

Một ví dụ đơn giản là việc phân loại học sinh hay phân loại việc bảo về luận án Tiến sỹ trên cơ sở cho điểm theo các tiêu chuẩn thành các mức đánh giá *suất sắc, khá*,... có thể xem là một phụ thuộc tương tự nếu các khoảng điểm phân định không ấn định tiền định, nghĩa là có thể thay đổi tùy theo yêu cầu quản lý hay tuyển dụng nhân viên của, chẳng hạn, các đơn vị doanh nghiệm.

Vì vậy ta giả thiết rằng, trong một bối cảch ứng dụng nhất định, dựa trên dữ liệu thực tế của một hệ CSDL, chúng ta có thể xác định được mức độ tương tự $k(A)$ ứng với mỗi thuộc tính ngôn ngữ A phù hợp với các phụ thuộc dữ liệu mờ. Nghĩa là, trên cùng một CSDL ngôn ngữ, có thể có những hệ phụ thuộc dữ liệu mờ khác nhau, tùy theo quan điểm khai thác dữ liệu của nhóm người sử dụng. Giả sử đối với một CSDL cho một ứng dụng, một cộng đồng người sử dụng chọn cho mỗi thuộc tính ngôn ngữ A một độ tương tự $k(A)$. Để tiện, ta ngầm định đối với các thuộc tính kinh điển $k(A) = 0$ và ký hiệu $\kappa = \{k(A) : A \in U\}$.

Với X là một tập con của U và t, s là hai bộ tùy ý trên U, ta nói hai bộ này là tương tự mức κ trên tập X, và viết $t[X] =_\kappa s[X]$, nếu với mọi $A \in X$, ta có $t[A] =_{k(A)} s[A]$.

Định nghĩa 3.1. Giả sử R là một lược đồ quan hệ của một CSDL ngôn ngữ, r là một quan hệ trên R và giả sử $=_{k(A)}$ là quan hệ tương tự mức $k(A)$ của thuộc tính A nếu A là thuộc tính ngôn ngữ, và nó là quan hệ đẳng thức thông thường nếu A là thuộc tính kinh điển. Xét hai tập con $X,\ Y \subseteq U$. Ta nói Y phụ thuộc hàm tương tự mức $k(U)$ vào X, ký hiệu là

$f = X_\kappa \to Y$, trong quan hệ r, nếu ta có:

$$\forall t, s \in r,\ t[X] =_\kappa s[X] \Rightarrow t[Y] =_\kappa s[Y].$$

Khi đó, ta cũng nói r tuân theo ràng buộc $X_\kappa\!\!\rightarrow Y$, hay ta cũng nói $X_\kappa\!\!\rightarrow Y$ thỏa trong quan hệ r.

Có thể thấy, theo định nghĩa trên, một phụ thuộc hàm (kinh điển) $X \to Y$ cũng là một phụ thuộc tương tự mức κ bất kỳ. Tuy nhiên điều ngược lại không đúng.

Ta xét một ví dụ về phụ thuộc tương tự.

Ví dụ 3.1. Để đơn giản ta xét lược đồ quan hệ R với 4 thuộc tính sau: 1 thuộc tính kinh điển Tên tổ chức nghiên cứu (TÊN TCNC), 3 thuộc tính ngôn ngữ là SỐ ĐIỂMCÔNG TRÌNH CÔNG BỐ TRONG NƯỚC (CTTN), SỐ ĐIỂM CÔNG TRÌNH CÔNG CÔNG BỐ NƯỚC NGOÀI (CTNN) và NĂNG LỰC TỔ CHỨC NGHIÊN CỨU (NLTC). Miền giá trị của thuộc tính HT và NLTC là các xâu ký tự. Miền giá trị thực của các thuộc tính CTTN là $[0, 60]$, miền giá trị thực của CTNN là $[0, 20]$ và miền thực của NLTC là thang điểm đánh giá trong đoạn [0, 10] và giá trị ngôn ngữ của NLTC được sinh từ các phần từ sinh của ĐSGT và sẽ được xác định sao cho nó tương đồng với nhãn ngôn ngữ suất sắc, giỏi, khá, trung bình, kém. Miền giá trị ngôn ngữ của các thuộc tính ngôn ngữ đều có cùng tập các xâu giống nhau với tập các phần tử sinh là $\{\mathbf{0}, \text{nhỏ}, \boldsymbol{W}, \text{lớn}, \mathbf{1}\}$ và tập các gia tử là $\{k', h', h, k\}$, trong đó k' là gia tử ít, h' là gia tử khá, h là gia tử hơn và k là gia tử rất. Chẳng hạn, xâu $h'c^-$ là từ khá nhỏ (chẳng hạn một số khá nhỏ công trình) còn hc^+ là từ lớn hơn (chẳng hạn một số lớn hơn cán bộ nghiên cứu),...

Mặc dù các thuộc tính ngôn ngữ đang xét có cùng tập các xâu, nhưng ngữ nghĩa định lượng của chúng khác nhau. Chẳng hạn, giả sử rằng nhìn chung các nhóm nghiên cứu có số công trình nghiên cứu khá lớn và ta mong muốn số lượng các nhóm xứng đáng nhận đánh giá suất sắc trên tiêu chí số công trình công bố trong nước không nhiều thì khoảng mờ tương ứng với thang đánh giá ngôn ngữ suất sắc là tương đối nhỏ. Nhưng giả định thực tế số nhóm có số lượng công trình công bố ở nước ngoài lớn hơn 12 là nhỏ thì khoảng mờ tương ứng với thang đánh giá suất sắc lại tương đối lớn. Với lý do đó ta chọn các tham số độ đo tính mờ ngôn ngữ như sau:

- Đối với thuộc tính CTTN: $fm(\text{lớn}) = fm(c^+) = 0,35$, $fm(\text{nhỏ}) = fm(c^-) = 0,65$, $\mu(\text{khá}) = 0,25$, $\mu(\text{ít}) = 0,20$, $\mu(\text{hơn}) = 0,15$ và $\mu(\text{rất}) = 0,40$. Như vậy, $\alpha = 0,45$ và $\beta = 0,55$.

Ta phân hoạch đoạn $[0, 60]$ thành 5 khoảng tương tự mức 1 là:

$fm(kc^+) \times 60 = 0,35 \times 0,35 \times 60 = 7,35$. Vậy, $S(\mathbf{1}) \times 60 = (52, 65, 60, 00]$.

$(fm(h'c^+) + fm(hc^+)) \times 60 = (0,25 \times 0,35 + 0,15 \times 0,35) \times 60 = 8,4$ và

$S(c^+) \times 60 = (44, 25, 52, 65]$.

$(fm(k'c^-) + fm(k'c^+)) \times 60 = (0,25 \times 0,65 + 0,25 \times 0,35) \times 60 = 15,0$ và

$S(\boldsymbol{W}) \times 60 = (29, 25, 44, 25]$.

$(fm(h'c^-) + fm(hc^-)) \times 60 = (0,25 \times 0,65 + 0,15 \times 0,65) \times 60 = 15,6$ và

$S(c^-) \times 60 = (13, 65, 29, 25]$.

Cuối cùng, $S(\mathbf{0}) \times 60 = [00, 00, 13, 65]$.

- Đối với thuộc tính CTNN: Do thực tế, các nhóm nghiân cứu còn ít công bố ở nước ngoài nên ta sẽ xác định tham số theo khuynh hướng "dễ dãi" đối với tiêu chí này: $fm(\text{lớn})$

$= fm(c^+) = 0,6$, fm(nhỏ) $= fm(c^-) = 0,4$, μ(khá) $= 0,15$, μ(ít) $= 0,25$, μ(hơn) $= 0,25$ và μ(rất) $= 0,35$. Như vậy, $\alpha = 0,6$ và $\beta = 0,4$.

Ta phân hoạch đoạn $[0, 20]$ thành 5 khoảng tương tự mức 1 là:

$fm(kc^+) \times 60 = 0,35 \times 0,6 \times 20 = 4,20$. Vậy, $S(\mathbf{1}) \times 60 = (15, 80, 20, 00]$.

$(fm(h'c^+) + fm(hc^+)) \times 20 = (0,25 \times 0,6 + 0,15 \times 0,6) \times 20 = 4,80$ và

$S(c^+) \times 60 = (11, 00, 15, 80]$.

$(fm(k'c^-) + fm(k'c^+)) \times 20 = (0,25 \times 0,6 + 0,25 \times 0,4) \times 20 = 5,00$ và

$S(\boldsymbol{W}) \times 60 = (06, 00, 11, 00]$.

$(fm(h'c^-) + fm(hc^-)) \times 20 = (0,25 \times 0,4 + 0,15 \times 0,4) \times 20 = 3,20$ và

$S(c^-) \times 60 = (02, 80, 06, 00]$.

Cuối cùng, $S(\mathbf{0}) \times 60 = [00, 00, 02, 80]$.

- Đối với thuộc tính NLTC: Ta giả thiết rằng việc đánh giá theo thủ tục là các chuyên gia cho điểm, tính trung bình cộng và sau đó phân theo thang bậc *suất sắc, giỏi, khá, trung bình, kém*. Thông thường, ta sử dụng các thang ngôn ngữ này tương ứng với các khoảng phân hoạch $[9, 5, 10, 0], [8, 5, 9, 5), [7, 0, 8, 5), [5, 0, 7, 0)$ và $[0, 0, 5, 0)$. Tuy nhiên, khi ta xác định các phương trình tính các khoảng tương tự như trên để xác định các tham số độ đo tính mờ thì hệ vô nghiệm. Lý do là khoảng $[0, 0, 5, 0]$ có độ dài khá bất thường. Để giải quyết vấn đề này, ta giữ nguyên các khoảng phân hoạch trên, trừ khoảng bất thường có độ dài bằng tổng độ dài các khoảng còn lại, được thay bằng khoảng $[l, 5, 0)$, với l là tham số, để tính các tham số độ đo tính mờ. Sau đó trên thực tế ta vẫn sử dụng khoảng $[0, 0, 5, 0)$. Với quy ước đó ta tính được các tham số sau: μ(ít) $= 0,25$, fm(lớn) $= fm(c^+) = 0,333$, fm(nhỏ) $= fm(c^-) = 0,667$, μ(khá) $+ \mu$(hơn) $= 0,5$ và μ(rất) $= 0,25$. Khi đó các khoảng phân hoạch trên được xem là các khoảng tương tự mức 1.

Quan hệ r trong ví dụ này được cho trong Bảng 1.

Bảng 1. Quan hệ với thuộc tính ngôn ngữ

TÊN TCNC	CTTN	CTNN	NLTC
#1	55	16	9,9
#2	12	2	4,4
#3	46	17	9,8
#4	46	17	9,5
#5	53	14	9,1
#6	56	13	8,8
#7	17	4	6,2
#8	23	8	7,9

Trong quan hệ này, TÊN TCNC là khóa (theo nghĩa kinh điển) nhưng tập hai thuộc tính CTTN và CTNN không xác định hàm thuộc tính NLTC do các dữ liệu trong hai hàng 3 và 4. Điều này trên thực tế có thể xảy ra do kết quả bỏ phiếu đánh giá. Căn cứ thực tiễn có thể do tập thể nghiên cứu #3 do một cán bộ khoa học trẻ có năng lực phụ trách và được các ủy viên hội đồng thưởng điểm trong đánh giá theo quan điểm của họ. Cũng tương tự, có thể giáo viên đánh giá học viên này suất sắc hơn học viên kia dù điểm bằng nhau hoặc kém hơn.

Tuy nhiên, ta có thể kiểm chứng rằng {CTTN, CTNN} $\underset{\kappa}{\rightarrow}$ NLTC, trong đó $k(U)$ đều là mức 1 trên các thuộc tính ngôn ngữ.

Ta cần nhấn mạnh rằng, phụ thuộc tương tự {CTTN, CTNN} $_\kappa\rightarrow$ NLTC là một ràng buộc mềm dẻo, thể hiện một quy định đánh giá buộc phải tuân thủ. Chẳng hạn, ở hàng 3 và 4, nếu tiêu chí số điểm CTNN được "phân loại" suất sắc (vì nó thuộc phân hoạch lớn nhất) còn đối với tiêu chí CTTN được "phân loại" giỏi (thuộc phân hoạch thứ hai) thì được đánh giá thuộc nhóm suất sắc. Tuy nhiên, nếu thuộc tính NLTC chỉ nhận các giá trị ngôn ngữ thì nó sẽ trở thành phụ thuộc hàm. Nhưng việc cho phép thuộc tính này nhận giá trị thực có ưu điểm là ta có thể thay đổi cách phân hoạch các khoảng điểm đánh giá, nghĩa là cho phép ngữ nghĩa của các giá trị ngôn ngữ không tiền định.

Như vậy, phụ thuộc tương tự, bên cạnh ngữ nghĩa mờ vẫn đề cập, nó còn có ý nghĩa thực tiễn quan trọng.

Gọi $\mathcal{F}_\kappa$ là họ tất cả các phụ thuộc hàm mức κ trên lược đồ CSDL DB. Ta ký hiệu $\mathcal{F}_\kappa^*$ là tập tất cả các phụ thuộc tương tự f mức κ mà là hệ quả ngữ nghĩa của $\mathcal{F}_\kappa$, tức là, với mọi quan hệ r trên U, nếu r thỏa các thụ thuộc dữ liệu trong $\mathcal{F}_\kappa$ thì r cũng thỏa f. Ta có thể dễ dàng kiểm chứng họ các phụ thuộc tương tự $\mathcal{F}_\kappa^*$ có các tính chất sau.

Định lý 3.1. *Trong CSDL ngôn ngữ với tập vũ trụ các thuộc tính U, họ $\mathcal{F}_\kappa^*$ thỏa mãn các tính chất sau:*

(i) Phản xạ: $X_\kappa \rightarrow X \in \mathcal{F}_\kappa^*$.
(ii) Gia tăng: $X_\kappa \rightarrow Y \in \mathcal{F}_\kappa^* \Rightarrow XZ_\kappa \rightarrow YZ \in \mathcal{F}_\kappa^*$.
(iii) Bắc cầu: $X_\kappa \rightarrow Y,\ Y_\kappa \rightarrow Z \in \mathcal{F}_\kappa^* \Rightarrow X_\kappa \rightarrow Z \in \mathcal{F}_\kappa^*$.

Cho tập $\mathcal{F}_\kappa$ và ký hiệu $\mathcal{F}_\kappa^+$ là tập nhỏ nhất chứa $\mathcal{F}_\kappa$ và đóng đối với các tính chất, được gọi là các tiên đề nêu trong Định lý 3.1. Khi đó, Định lý 3.1 bảo đảm rằng nếu quan hệ r thỏa $\mathcal{F}_\kappa$ thì nó cũng thỏa $\mathcal{F}_\kappa^+$, hay $\mathcal{F}_\kappa^+ \subseteq \mathcal{F}_\kappa^*$.

Định lý 3.2. *Hệ tiên đề $(i)-(iii)$ trong Định lý 3.1 là đầy đủ, nghĩa là $\mathcal{F}_\kappa^+ = \mathcal{F}_\kappa^*$ hay ta nói hệ tiên đề (i) - (iii) là đủ để sinh ra toàn bộ tập $\mathcal{F}_\kappa^*$.*

Chứng minh. Ta dễ dàng thấy rằng trong một quan hệ 2-bộ r với các bộ t và s chỉ chứa các giá trị lớn nhất và nhỏ nhất trong miền giá trị thuộc tính, điều kiện $t[X] =_\kappa s[X]$ là tương đương với điều kiện $t[X] = s[X]$, vì khi đó các khoảng phân hoạch (do nó chứa ít nhất hai lớp tương đương) của các thuộc tính ngôn ngữ chỉ chứa một trong hai giá trị này. Do đó, nếu quan hệ 2-bộ r như vậy thỏa một phụ thuộc tương tự nào đó (mức κ), thì nó cũng thỏa phụ thuộc đó khi được xem như là một phụ thuộc hàm. Vì vậy, cũng như đối với họ phụ thuộc hàm kinh điển, nếu $f = X_\kappa\rightarrow X \notin \mathcal{F}_\kappa^+$ thì tồn tại một quan hệ 2-bộ r sao cho r thỏa $\mathcal{F}_\kappa$ nhưng không thỏa phụ thuộc tương tự f. Nghĩa là, hệ tiên đề trong Định lý 3.1 là đầy đủ.

Như vậy, ta thấy ở mức ký pháp, phụ thuộc tương tự trùng với phụ thuộc hàm kinh điển nhưng ngữ nghĩa khác nhau, đặc biệt một quan hệ r có thể thỏa một quan hệ hàm tương tự f nhưng nó không nhất thiết thỏa f với tư cách là một phụ thuộc hàm kinh điển. Vì vậy, trong một CSDL ngôn ngữ có thể tồn tại hai loại ràng buộc dữ liệu: (i) Phụ thuộc hàm kinh điển, nó có giá trị cho thiết kế CSDL theo các dạng chuẩn. (ii) Phụ thuộc hàm tương tự nhưng quan hệ r không nhất thiết thỏa nó như là một phụ thuộc hàm. Tuy nhiên, nó vẫn là một ràng buộc yếu đối với CSDL như ta thấy trong Ví dụ 3.1. Do đó, một CSDL ngôn ngữ có thể biểu thị bằng một bộ sau

$$DB = \{U; \mathcal{F},\ \mathcal{F}_\kappa\},$$

trong đó $\mathcal{F}$ là tập phụ thuộc hàm, còn $\mathcal{F}_\kappa$ là tập phụ thuộc tương tự và $\mathcal{F} \cap \mathcal{F}_\kappa = \emptyset$.

Như thường lệ, Y phụ thuộc hàm vào X sẽ được ký hiệu bằng $X \to Y$ và lưu ý rằng đối với phụ thuộc hàm như vậy X hay Y vẫn chứa các thuộc tính ngôn ngữ, tuy nhiên nó xem các giá trị ngôn ngữ như là các ký hiệu (mức cú pháp).

Phụ thuộc hàm là một ràng buộc của cơ sở dữ liệu ở mức cú pháp, tức là mức ký hiệu, phục vụ cho việc thao tác dữ liệu ở mức ký hiệu. Ví dụ, nếu $f = X \to A$ với A là thuộc tính ngôn ngữ LỨA TUỔI, thì giá trị "trẻ" xuất hiện trong cột thuộc tính A cũng bị ràng buộc bởi phụ thuộc hàm f, trong khi có thể có phụ thuộc tương tự $Y_\kappa \to A$. Phụ thuộc tương tự trong $\mathcal{F}_\kappa$ cũng là một ràng buộc CSDL nhưng không chặt vì nó cho phép quan hệ giữa các giá trị thuộc tính không nhất thiết tuân theo phụ thuộc hàm. Nó xác định mối quan hệ giữa các giá trị của một thuộc tính trong một phân hoạch độ tương tự mức k nào đó.

Do hai loại phụ thuộc hàm này có quan hệ gần gũi với nhau, cả ở mức hình thức hóa, nên chúng có thể sinh thêm các phụ thuộc hàm mới không nằm trong $\mathcal{F}^+ \cup \mathcal{F}_\kappa^+$.

Gọi $\mathcal{F}_{DB}$ là tập các phụ thuộc hàm được sinh ngữ nghĩa từ $\mathcal{F}$ và $\mathcal{F}_\kappa$ theo nghĩa $\mathcal{F}_{DB}$ là tập nhỏ nhất sao cho nó chứa mọi phụ thuộc $X \to Y$ thỏa trong mọi quan hệ r trên U mà nó bị ràng buộc bởi các phụ thuộc dữ liệu trong $\mathcal{F} \cup \mathcal{F}_\kappa$. Tương tự như vậy, ta định nghĩa tập $\mathcal{F}_{DB}$ là tập tất cả các phụ thuộc tương tự được sinh ngữ nghĩa từ hai tập $\mathcal{F}$ và $\mathcal{F}_\kappa$.

Trong cách tiếp cận đại số, ta có thể thiết lập mối liên hệ khá chỉnh giữa hai loại phụ thuộc dữ liệu này, cụ thể nó được phát biểu trong mệnh đề sau.

Định lý 3.3. *Trong CSDL ngôn ngữ $DB = \{U; \mathcal{F}, \mathcal{F}_\kappa\}$ ta có:*

(i) $X \to Y \in \mathcal{F}_{DB}$ và $Y_\kappa \to Z \in \mathcal{F}_{DB} \Rightarrow X_\kappa \to Z \in \mathcal{F}_{DB}$, nếu X không chứa thuộc tính ngôn ngữ.

(ii) $X_\kappa \to Y \in \mathcal{F}_{DB} \Rightarrow X \to Y \in \mathcal{F}_{DB}$, nếu Y không chứa thuộc tính ngôn ngữ.

(iii) $X_\kappa \to Y \in \mathcal{F}_{DB}$ và $Y \to Z \in \mathcal{F}_{DB} \Rightarrow X \to Z \in \mathcal{F}_{DB}$, nếu Y không chứa thuộc tính ngôn ngữ.

(iv) $X_\kappa \to Y \in \mathcal{F}_{DB}$ và $Z \to W \in \mathcal{F}_{DB} \Rightarrow XZ_\kappa \to YW \in \mathcal{F}_{DB}$, nếu Z không chứa thuộc tính ngôn ngữ.

Chứng minh. Việc chứng minh định lý không khó, nhưng cần diễn đạt cụ thể để phân biệt vai trò của hai loại phụ thuộc hàm đã đề cập. Trong các lập luận dưới đây, ta giả thiết quan hệ r thỏa các phụ thuộc dữ liệu trong $\mathcal{F}$ và $\mathcal{F}_\kappa$, t và s là hai bộ trong r.

Để chứng minh (i), ta giả sử $t[X] =_\kappa s[X]$. Vì X không chứa thuộc tính ngôn ngữ nên ta suy ra $t[X] = s[X]$. Do đó, $t[Y] = s[Y]$ ở mức ký hiệu và việc này kéo theo $t[Y] =_\kappa s[Y]$. Từ đó ta có $t[Z] =_\kappa s[Z]$, nghĩa là $X_\kappa \to Z$ thỏa trong r (hay $X_\kappa \to Z \in \mathcal{F}_{DB}$).

Khẳng định (ii) rút ra từ sự kiện là $t[X] = s[X]$ kéo theo $t[X] =_\kappa s[X]$, và do đó $t[Y] =_\kappa s[Y]$. Vì Y chỉ chứa các thuộc tính kinh điển nên biểu thức cuối cùng có nghĩa $t[Y] = s[Y]$, tức là $X \to Y$ thỏa trong r.

Để chứng minh (iii), giả sử $t[X] =_\kappa s[X]$, khi đó ta có $t[Y] =_\kappa s[Y]$. Tương tự như trên, vì Y không chứa thuộc tính ngôn ngữ, nên ta suy ra $t[Y] = s[Y]$ và do đó, theo giả thiết $Y \to Z \in \mathcal{F}$, ta thu được $t[Z] = s[Z]$ ở mức ký hiệu. Vậy, $X \to Z$ thỏa trong r.

Bây giờ ta chứng minh (iv). Theo tính chất (i), ta có $Z_\kappa \to W$ đúng trong quan hệ đang xét r. Vì, theo Định lý 3.1, các phụ thuộc hàm tương tự trong quan hệ r đóng đối với các tiên đề (i) - (iii), nên theo Tiên đề (ii) và từ giả thiết của (iv) ta có $XZ_\kappa \to YZ$ và $YZ_\kappa \to YZ$. Vậy, theo tiên đề bắc cầu (iii) ta thu được biểu thức cần chứng minh.

Lưu ý: Mặc dù khái niệm phụ thuộc hàm tương tự là sự mở rộng khá trực tiếp từ khái niệm phụ thuộc hàm (khi mà tất cả các thuộc tính đều là kinh điển), nhưng trong CSDL ngôn ngữ (và cả trong CSDL mờ), chúng ta cần phải phân biệt rõ ràng hai lớp phụ thuộc dữ liệu này vì vai trò của chúng rất khác nhau. Vì vậy, các giả thiết chứa cụm từ "không chứa thuộc tính ngôn ngữ" trong Định lý 3.3 là thiết yếu.

4. KẾT LUẬN

Cho đến nay việc nghiên cứu CSDL với thông tin mờ, không chắc chắn chủ yếu dựa trên cách tiếp cận của lý thuyết tập mờ và lý thuyết khả năng. ĐSGT cho ta một cách tiếp cận mới đến việc biểu diễn ngữ nghĩa của các khái niệm mờ, nó giải quyết các mối quan hệ ngữ nghĩa và như vậy nó chứa đựng các thông tin ngữ nghĩa ở mức tổng thể, mức hệ thống hơn. Vì vậy, mặc dù các khoảng mờ biểu thị lân cận về độ tương tự có vẻ giống như cách tiếp cận dựa trên giá trị khoảng, nhưng các lân cận trong cách tiếp cận ĐSGT không được phép tùy tiện mà chúng phải được xác định dựa trên các tham số độ đo tính mờ của ngôn ngữ. Nhờ các khoảng lân cận như vậy và nhờ ánh xạ định lượng với tham số là độ đo tính mờ của ngôn ngữ, các giá trị ngôn ngữ có giá trị thực trong miền tham chiếu làm đại diện cùng với hệ lân cận ngữ nghĩa đã cho phép ta chuyển các thao tác dữ liệu trong CSDL ngôn ngữ về các thao tác dữ liệu kinh điển làm cho việc tổ chức thao tác trở nên đơn giản và gần gũi hơn so với các cách tiếp cận khác.

Với ưu điểm như vậy, việc nghiên cứu phụ thuộc hàm mờ trong CSDL ngôn ngữ, được gọi là phụ thuộc hàm tương tự, trong ngữ cảnh các phụ thuộc kinh điển đã giải quyết. Hiện nay, vấn đề này vẫn còn mới mẻ và chưa được làm rõ. Do đó, trong bài báo này, tác giả đã làm rõ hơn mối quan hệ giữa phụ thuộc dữ liệu kinh điển và phụ thuộc mờ, và theo đó, một số vấn đề nảy sinh vẫn còn để ngỏ:

• Trong phạm vi nghiên cứu của bài báo này, giả thiết rằng hai tập $\mathcal{F}$ và $\mathcal{F}_\kappa$ là rời nhau và có vai trò hoàn toàn khác nhau, trong đó một phụ thuộc trong $\mathcal{F}_\kappa$ (chứ không phải trong $\mathcal{F}_{DB}$) không phải là phụ thuộc hàm. Về trực quan ta mong muốn $\mathcal{F}^+$ là tập tất cả các ràng buộc phụ thuộc hàm của CSDL $\mathcal{DB}$, và như vậy vấn đề đặt ra là liệu $\mathcal{F}^+ = \mathcal{F}_{DB}$?

• Tập $\mathcal{F}_{DB}$ thực sự lớn hơn tập $\mathcal{F}_\kappa^+$ và nếu câu trả lời cho câu hỏi trên là không đúng thì vấn đề tìm một hệ tiên đề đầy đủ cho hai tập $\mathcal{F}_{DB}$ và $\mathcal{F}_{DB}$ vẫn còn là để mở.

TÀI LIỆU THAM KHẢO

[1] T. K. Bhattacharjee and A. K. Mazumdar, Axiomatisation of fuzzy multivalued dependencies in fuzzy relational data model, *Fuzzy Sets and Systems* **96** (1998) 343–352.

[2] D. A. Chiang, L. R. Chow, and N. C. Hsien, Fuzzy information in extended fuzzy relational databases, *Fuzzy Sets and Systems* **92** (1997) 1–20.

[3] N. C. Ho, A model of relational databases with linguistic data of hedge algebras - based semantics, Hội thảo quốc gia lần thứ ba về "Nghiên cứu phát triển và ứng dụng CNTT và Truyền thông" ICT.rda'2006, 20-21/05/2006.

[4] N. C. Ho, Fuzziness in structure of linguistic truth values: a foundation for development of fuzzy reasoning, *Proc. of Int. Symp. on Multiple-Valued Logic*, Boston University, Boston, Massachusetts, IEEE Computer Society Press, May 26-28, 1987 (325–335).

[5] N. C. Ho, Quantifying hedge algebras and interpolation methods in approximate reasoning, *Proc. of the 5th Inter. Conf. on Fuzzy Information Processing*, Beijing, March 1-4, 2003 (105–112).

[6] N. C. Ho, N. V. Long, Complete and linear hedge algebras, fuzziness measure of vague concepts and linguistic hedges and application, (Best paper Award of the Conference), *AIP Conf. Proceed. on Computing Anticipatory Systems, CASYS'05*, Liege, Belgium, 8-13 August 2005 (ed. Daniel M. Dubois, 331-339).

[7] N. C. Ho, N. V. Long, Fuzziness measure on complete hedge algebras and quantitative semantics of terms in linear hedge algebras, *Fuzzy Sets and Systems* **158** (2007) 452–471.

[8] N. C. Ho, L. H. Chau, Quantitative semantics in hedge algebras and interpolation methods, *Proc. of ICT*, Hanoi (2003).

[9] N. C. Ho, H. V. Nam, Ordered structure-based semantics of linguistic terms of linguistic variables and approximate reasoning, *AIP Conf. Proceed. on Computing Anticipatory Systems, CASYS'99, 3th Inter. Conf.*, 1999 (98–116).

[10] N. C. Ho, H. V. Nam, A theory of refinement structure of hedge algebras and its application to linguistic-valued fuzzy logic. In D. Niwinski & M. Zawadowski (Eds), *Logic, Algebra and Computer Science***Vol. 46** (1999) (Banach Center Publications, Polish Scientific Publishers - PSP).

[11] N. C. Ho, H. V. Nam, Towards an algebraic foundation for a zadeh fuzzy logic, *Fuzzy Set and System* **129** (2002) 229–254.

[12] N. C. Ho, H. V. Nam, T. D. Khang, and L. H. Chau, Hedge Algebras, Linguistic- valued Logic and their Application to Fuzzy Reasoning, *Inter. J. of Uncertainty, Fuzziness and Knowledge-Based System* **7** (1999) 347–361.

[13] N. C. Ho and W. Wechler, Hedge algebras: An algebraic approach to structures of sets of linguistic domains of linguistic truth variable, *Fuzzy Sets and Systems* **35** (1990) 281–293.

[14] N. C. Ho and W. Wechler, Extended hedge algebras and their application to fuzzy logic, *Fuzzy Sets and Systems* **52** (1992) 259–281.

[15] N. C. Ho, Quantifying hedge algebras and interpolation methods in approximate reasoning, *Proc. of the 5th Inter. Conf. on Fuzzy Information Processing*, Beijing, March 1-4, 2003 (105–112).

[16] N. C. Ho, L. H. Chau, T. D. Khang, and H. V. Nam, Hedge algebras, linguistic- valued logic and their application to fuzzy reasoning, *International Journal of Uncertainty, Fuzziness and Knowledge-Based System* **7** (1999) 347–361.

[17] S. Jyothi, M. Syam Babu, Multidependencies in fuzzy relational databases and lossless join decomposition, *Fuzzy Sets and Systems* **88** (1997) 315–332.

[18] L. Di Lascio, A. Gisolfi, and V. Loia, A new model for linguistic modifiers, *International Journal of Approximate Reasoning* **15** (1996) 25–47.

[19] L. Di Lascio and A. Gisolfi, Averaging linguistic truth values in fuzzy approximate reasoning, *International Journal of Intelligent Systems* **13** (1998) 301–318.

[20] Weip Yi Liu, A relational data model with fuzzy inheritance dependencies, *Fuzzy Sets ans Systems* **89** (1997) 205–213.

[21] M. Umano and O. Freedom, A fuzzy database system, *Fuzzy Information and Decision Processes,* North-Holland, Armsterdam, 1982 (339–347).

Nhận bài ngày 08 - 5 - 2007
Nhận lại sau sửa ngày 19 - 7 -2007