Tạp chí Tin học và Điều khiển học, T.25, S.4 (2009), 314–332

THỐNG NHẤT DỮ LIỆU VÀ XÂY DỰNG QUAN HỆ TƯƠNG TỰ TRONG CƠ SỞ DỮ LIỆU NGÔN NGỮ BẰNG ĐẠI SỐ GIA TỬ*

NGUYỄN CÁT HỒ[1], LÊ XUÂN VINH[2], NGUYỄN CÔNG HÀO[3]

[1] *Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam*
[2] *Trường Đại học Quy Nhơn*
[3] *Trường Đại học khoa học Huế*

Abstract. Fuzzy databases under several different approaches allow not only traditional crisp data and fuzzy sets representing the meaning of vague linguistic values, but also data in the form of finite subsets of attribute domains and, especially, null values of different types such as "inapplicable", "missing" or "at present unknown", and so on. In this paper, the semantics of linguistic terms in fuzzy databases will be considered based on hedge algebra structure and, consequently, the databases under such data semantics are called linguistic databases. For these databases, a unified method of fuzzy data representation in which each datum of all types under consideration is represented by a set of intervals in the respectice attribute domains, and therefore it is a method of interval representation. Note that these intervals are defined relying on the neighbourhoods of a semantics-based topology on attribute domains. In these databases, the concept of fuzzy equality and matching relations of degee k between data of different types, where k is a positive interger indicating the length of terms in the respective hedge algebra, are proposed and studies. Based on the interval representation of the data in databases, fuzzy queries to the linguistic databases will be transformed in to (crisp) ordinary ones and, hence, a great benefit is obtained.

Tóm tắt. Cơ sở dữ liệu mờ, với những cách tiếp cận khác nhau, ngoài dữ liệu rõ truyền thống và tập mờ biểu thị giá trị ngôn ngữ mờ còn cho phép chứa những dữ liệu dạng tập con hữu hạn của miền thuộc tính, đặc biệt các dạng dữ liệu "null" như dữ liệu không xác định, dữ liệu "at present unknown",...Trong bài báo, chúng tôi trình bày một phương pháp biểu diễn các dạng dữ liệu này một cách thống nhất bằng tập các khoảng, trong đó dữ liệu ngôn ngữ mờ được biểu thị bằng tập các khoảng được xác định từ lân cận tôpô ngữ nghĩa trên miền trị của thuộc tính. Vì vậy nó được gọi là phương pháp biểu diễn khoảng. Các khái niệm bằng nhau mờ, quan hệ đối sánh mờ mức k với k là độ dài của biểu diễn chính tắc của các giá trị ngôn ngữ trong đại số gia tử thích hợp được giới thiệu và nghiên cứu. Trên cơ sở đó một truy vấn ngôn ngữ mờ sẽ chuyển đổi được về một truy vấn thông thường khá tiện lợi cho việc xử lý.

1. GIỚI THIỆU

Đối với cơ sở dữ liệu mờ chứa giá trị ngôn ngữ, cách tiếp cận kinh điển để biểu diễn các

*Nghiên cứu được thực hiện dưới sự tài trợ của Quỹ phát triển Khoa học - Công nghệ Quốc gia Nafosted

giá trị này là sử dụng tập mờ [20, 21,16]. Chẳng hạn đối với thuộc tính A, một giá trị ngôn ngữ mờ của thuộc tính này được biểu diễn bởi tập mờ $A(x)$ trên không gian tham chiếu của A. Để so sánh độ tương tự giữa A và B, một số độ đo $Com(A, B)$ đo độ tương thích giữa hai tập mờ A và B được định nghĩa dựa trên những đặc trưng của các dạng tập mờ tam giác, tập mờ hình thang [16]. Một cách tiếp cận khác là dựa trên lý thuyết khả năng: sử dụng một phân phối khả năng $\pi_A : D_A \to [0, 1]$ với $\pi_A(d)$ biểu thị khả năng thuộc tính A nhận giá trị d. Dựa trên lý thuyết này, có thể định nghĩa độ đo $Poss(A, B)$ - khả năng bằng nhau giữa A và B; độ đo $Nec(A, B)$ biểu thị mức độ của sự kiện "thông tin có trong B bị bao hàm trong A" và có thể định nghĩa nhiều phép toán khác. Một cách tiếp cận nữa là sử dụng đại số gia tử (ĐSGT), một mô hình toán học về cấu trúc thứ tự ngữ nghĩa của miền giá trị của biến ngôn ngữ. Mỗi giá trị ngôn ngữ được biểu diễn bởi một phần tử trong một ĐSGT thích hợp. Cách biểu diễn này cho phép thực hiện những thao tác trực tiếp trên ngôn ngữ, bảo toàn được ngữ nghĩa của dữ liệu ngôn ngữ theo quan điểm tính toán trên các từ [24]. Tuy nhiên, do tính chất định tính của dữ liệu ngôn ngữ nên khó biểu diễn thống nhất với các dạng dữ liệu khác như số, khoảng và đặc biệt hơn là tập các giá trị rời rạc, "unknown", "missing", "inapplicable" [8]. Vì vậy cần phải có phương pháp định lượng ngữ nghĩa các giá trị ngôn ngữ làm cơ sở để biểu diễn một cách thống nhất các dạng dữ liệu trên.

Về cơ bản phương pháp này dựa trên ý tưởng chính như sau. Miền giá trị ngôn ngữ của mỗi thuộc tính được nhúng vào một ĐSGT phù hợp. Sử dụng cách lượng hóa của ĐSGT, mỗi giá trị ngôn ngữ x sẽ xác định một khoảng $\mathfrak{I}(x)$ trên miền trị của thuộc tính tương ứng. Kích thước của $\mathfrak{I}(x)$ phụ thuộc vào độ dài biểu diễn chính tắc của x trong ĐSGT. Tập tất cả các khoảng $\mathfrak{I}(hx)$ với h là gia tử tùy ý sẽ phân hoạch khoảng $\mathfrak{I}(x)$. Nếu xem chúng được sinh ra từ $\mathfrak{I}(x)$ thì tập tất cả các khoảng biểu diễn cho tất cả các giá trị của ĐSGT sẽ là một cây. Mỗi giá trị thực trong một khoảng $\mathfrak{I}(x)$ ở độ sâu k được xem là tương thích ngữ nghĩa mức k với x. Trên cơ sở này, chúng ta sẽ xây dựng phương pháp để biểu diễn thống nhất các dạng dữ liệu nói trên. Hơn nữa, tập các khoảng ở độ sâu k sẽ phân hoạch miền trị của thuộc tính. Chúng ta biết rằng mỗi phân hoạch xác định một quan hệ tương đương. Các phần tử thuộc cùng lớp tương đương sẽ được xem là bằng nhau mức k. Từ đó chúng ta có thể xây dựng các quan hệ đối sánh và chuyển đổi các truy vấn mờ trên các cơ sở dữ liệu ngôn ngữ thành các truy vấn thông thường để xử lý.

Bài báo gồm 5 mục. Mục 2 nhắc lại những kết quả cơ bản về phương pháp lượng hóa trong ĐSGT; Mục 3 dành cho việc trình bày quan điểm biểu diễn các dạng dữ liệu không đồng nhất; Mục 4 trình bày cơ sở để xây dựng các quan hệ tương tự, quan hệ đối sánh; Mục 5 là một số nhận xét kết luận cho bài báo.

2. CƠ SỞ LƯỢNG HÓA NGỮ NGHĨA CÁC GIÁ TRỊ NGÔN NGỮ

2.1. Một số khái niệm và tính chất cơ bản của ĐSGT

Như trên đã nói, ĐSGT là một trong những cách tiếp cận để phát hiện cấu trúc đại số của miền giá trị của biến ngôn ngữ. Theo quan điểm đại số, miền giá trị của biến ngôn ngữ $\mathfrak{X}$

được sinh ra từ tập các phần tử sinh thường gồm hai phần tử $G = \{c^+, c^-\}$ bởi sự tác động của các gia tử trong H. Quan hệ thứ tự $\leq$ trên cấu trúc này được xác lập từ ngữ nghĩa các từ ngôn ngữ. Cấu trúc thứ tự cảm sinh trực tiếp như vậy chính là điểm khác biệt so với các cách tiếp cận khác. Thêm một số phần tử, toán tử đặc biệt, ĐSGT là một đại số trừu tượng $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$, trong đó $C = \{0, W, 1\}$ là các hằng; Φ, Σ là các toán tử lấy giới hạn của tập các phần tử được sinh ra khi chịu tác tác động liên tiếp của các gia tử trong H. Một cách khác, nếu kí hiệu $H(x) = \{h \ldots h'x | h, \ldots, h' \in H\}$ thì $\Phi x = infimumH(x)$ và $\Sigma x = supremumH(x)$. Như vậy, ĐSGT $\underline{AX}$ được xây dựng trên nền của một ĐSGT $AX = (X, G, C, H, \leq)$, ở đây $X = H(G)$, bằng cách bổ sung hai toán tử Φ, Σ. Khi đó $X = X \cup Lim(G)$ với $Lim(G)$ là tập các phần tử giới hạn: với mọi $x \in Lim(G)$, tồn tại $u \in X$ sao cho $x = \Phi u$ hoặc $x = \Sigma u$. Các phần tử giới hạn này được bổ sung vào ĐSGT AX để làm cho các phép tính mới có nghĩa và vì vậy $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ được gọi là ĐSGT đầy đủ [11]. Còn lại, những phần tử $x \in \underline{X} - Lim(G)$ đều có thể biểu diễn được ở dạng $x = h_n ... h_1 c$, với $c \in G$. Nếu $h_i h_{i-1} ... h_1 c \neq h_{i-1} ... h_1 c$, với mọi $1 < i \leq n$ thì $h_n ... h_1 c$ được gọi là dạng biểu diễn chính tắc của x. Dạng biểu diễn chính tắc của mỗi phần tử $x \in X$ là duy nhất. Vì vậy ta có thể định nghĩa độ dài của x, kí hiệu là $|x|$, bằng tổng số các gia tử và phần tử sinh có mặt trong biểu diễn chính tắc của nó.

Trong H có những gia tử khi tác động thì có khuynh hướng làm mạnh lên ngữ nghĩa của phần tử sinh nguyên thủy gọi là *các gia tử dương*, có gia tử thì làm yếu đi gọi là *các gia tử âm*. Chẳng hạn, xem $True$ là phần tử sinh nguyên thủy của biến ngôn ngữ $Truth$ thì $Very$ là gia tử dương vì $True \leq VeryTrue$, và $Little$ là gia tử âm vì $LittleTrue \leq True$. Như vậy, H có thể được chia thành hai tập rời nhau H^+ các gia tử dương và H^- các gia tử âm. Hơn nữa, hai gia tử trong trong mỗi tập H^+ (hay H^-) có thể sánh được hoặc không sánh được về khả năng tác động của nó đối với một giá trị ngôn ngữ. Nói chung, H^+ , H^- là các tập sắp thứ tự bộ phận. Nếu như H^+ , H^- là các tập sắp thứ tự toàn phần và $G \cup C = \{0, c^-, W, c^+, 1\}$ cũng vậy thì $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ được gọi là ĐSGT tuyến tính đầy đủ [12]; hơn nữa, nếu $hx \neq x$ với mọi $x \in X$ ngoại trừ 0, W, 1 thì nó được gọi là tự do.

So với những cách biểu diễn khác như tập mờ, cách biểu diễn bằng các phần tử trong ĐSGT cho phép so sánh các giá trị ngôn ngữ một cách dễ dàng hơn. Giả sử $x = h_n ... h_1 u$ là một biểu diễn của x đối với u, kí hiệu $x_{u|j}$ dùng để chỉ cho $h_{j-1} ... h_1 u$ với quy ước $x_{u|1} = u$. Để so sánh hai phần tử x và y ta dùng định lý sau đây.

Định lý 2.1. ([12, 13, 14]) *Cho $x = h_n ... h_1 u$ và $y = k_m ... k_1 u$ là hai biểu diễn chính tắc tương ứng của x và y đối với u. Khi đó, tồn tại chỉ số $j \leq \min\{m, n\} + 1$ sao cho $h_{j'} = k_{j'}$, với mọi $j' < j$ và*
(1) $x = y$ khi và chỉ khi $m = n$ và $h_j x_{u|j} = k_j x_{u|j}$;
(2) $x < y$ khi và chỉ khi $h_j x_{u|j} < k_j x_{u|j}$

2.2. Định lượng ngữ nghĩa các giá trị ngôn ngữ

Theo trực giác, mỗi giá trị ngôn ngữ có một độ mờ nhất định tùy thuộc vào mức độ xác

định của thông tin mà nó mang lại. Chẳng hạn, *rất-trẻ* có mức độ xác định cao hơn *trẻ* nên độ mờ của nó thấp hơn độ mờ của *trẻ*. Tổng quát, độ mờ của hx thấp hơn độ mờ của x và dễ thấy bao giờ ta cũng có $H(hx) \subseteq H(x)$. Nói một cách khác có những tập các giá trị ngôn ngữ có kích thước khác nhau chứa x và các tập nhỏ hơn chứa các giá trị ngôn ngữ có ngữ nghĩa gần ngữ nghĩa của x hơn. Điều này gợi ý cho chúng ta xem $H(x)$ như là một lân cận tôpô dựa trên ngữ nghĩa của x. Họ $\mathfrak{V} = \{H(x)|x \in H(G)\}$ thỏa các điều kiện của một cơ sở tôpô nên nó được gọi là *cơ sở tôpô ngữ nghĩa* $\mathfrak{T}$ trên $X = H(G)$ với một số tính chất cơ bản như sau:

a) $H(hx) \subseteq H(x)$, với mọi $h \in H$ và $x \in X$.

b) $H(hx) \cap H(kx) = \emptyset$, nếu $h, k \in H$ và $hx \neq kx$; nghĩa là lân cận của các phần tử có độ dài bằng nhau sẽ rời nhau.

c) $H(x) = \bigcup_{h \in H \cup I} H(hx)$, ở đây toán tử I được định nghĩa là $Ix = x$, $\forall x \in X$; nghĩa là lân cận của tất cả các phần tử có dạng hx là một phân hoạch trên lân cận của x.

Như vậy, tính mờ của x liên quan đến kích thước của $H(x)$. Giả sử $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ là một ĐSGT tuyến tính đầy đủ, tự do của biến ngôn ngữ $\mathfrak{X}$ và f là một ánh xạ liên tục bảo toàn thứ tự từ $X \to [0,1]$. Khi đó, độ dài của khoảng con bé nhất $\mathfrak{I}(x) \subseteq [0,1]$, chứa $f(H(x))$ có thể đặc trưng cho độ mờ của x. Vì vậy, ta có một hàm $fm : X \to [0,1]$ xác định độ mờ cho các phần tử trong X bằng cách đặt $fm(x) = |\mathfrak{I}(x)|$; và dựa vào đó độ mờ của các gia tử $\mu(h)$, $h \in H$ cũng được xác định với một số tính chất được phát biểu qua mệnh đề sau. Trong mệnh đề này giả sử rằng $H^+ = \{h_1, ..., h_p\}$, $H^- = \{h_{-1}, ..., h_{-q}\}$, ở đây $h_1 < ... < h_p$ và $h_{-1} < ... < h_{-q}$, với các số nguyên dương $p, q \geq 2$.

Mệnh đề 2.1. [12] *Cho fm là độ đo tính mờ trên X và $\mu(h)$ là độ mờ của các gia tử h, ta có:*

1) $fm(hx) = \mu(h) fm(x)$, $\forall x \in X$ và $fm(x) = 0$, $\forall x \in Lim(\underline{X})$;

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

3) $\sum_{-q \leq i \leq p, i \neq 0} fm(h_i c) = fm(c)$, với $c \in \{c^-, c^+\}$;

4) $\sum_{-q \leq i \leq p, i \neq 0} fm(h_i x) = fm(x)$, với $x \in \underline{X}$;

5) $\sum_{-q \leq i \leq -1} fm(h_i) = \alpha$ và $\sum_{1 \leq i \leq p} fm(h_i) = \beta$, ở đây $\alpha, \beta > 0$ và $\alpha + \beta = 1$.

Từ mệnh đề trên, ta thấy tập các khoảng mờ $\{\mathfrak{I}(h_i x)|i \in [-q^\wedge p]\}$ là một phân hoạch trên $\mathfrak{I}(x)$, ở đây $[-q^\wedge p]$ kí hiệu cho tập số nguyên $\{i| -q \leq i \leq p, i \neq 0\}$. Hơn nữa, chúng xếp trên $\mathfrak{I}(x)$ theo thứ tự sau đây:

Nếu $Sign(h_p x) = +1$, tức là $h_p x \geq x$ thì

$$\mathfrak{I}(h_{-q}x) \leq \mathfrak{I}(h_{-q+1}x) \leq ... \leq \mathfrak{I}(h_{-1}x) \leq \mathfrak{I}(h_1 x) \leq \mathfrak{I}(h_2 x) \leq ... \leq \mathfrak{I}(h_p x) \tag{2.1}$$

và ngược lại, nếu $Sign(h_p x) = -1$, tức là $h_p x \leq x$ thì

$$\mathfrak{I}(h_p x) \leq \mathfrak{I}(h_{p-1}x) \leq ... \leq \mathfrak{I}(h_1 x) \leq \mathfrak{I}(h_{-1}x) \leq \mathfrak{I}(h_{-2}x) \leq ... \leq \mathfrak{I}(h_{-q}x) \tag{2.2}$$

Một cách đầy đủ, gọi X_k là tập các phần tử có độ dài k, các khoảng mờ liên quan đến các phần tử có cùng độ dài có một số tính chất được phát biểu qua mệnh đề sau.

Mệnh đề 2.2. *Cho* $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ *là một ĐSGT tuyến tính đầy đủ và tự do. Khi đó, ta có*

(i) Với mỗi $x \in X_k$, *tập các khoảng mờ* $\{\mathfrak{I}(h_i x) | i \in [-q^\wedge p]\}$, *là một phân hoạch của* $\mathfrak{I}(x)$. *Hơn nữa,* $\mathfrak{I}(h_i x) \leq \mathfrak{I}(h_j x)$ *khi và chỉ khi* $h_i x \leq h_j x$, *với mọi* $i, j \in [-q^\wedge p]$, *ở đây* $U \leq Z$ *nghĩa là* $a \leq b, \forall a \in U$ *và* $\forall b \in Z$;

(ii) Tập các khoảng mờ $I_k = \{\mathfrak{I}(x) | x \in X_k\}$ *là một phân hoạch trên [0,1], đồng thời* I_k *và* X_k *đẳng cấu bảo toàn thứ tự tuyến tính;*

(iii) I_{k+1} *mịn hơn* I_k, *tức là mỗi khoảng mờ thuộc* I_{k+1} *bao giờ cũng nằm trong một khoảng mờ thuộc* I_k. *Cụ thể là mỗi* $x \in X_{k+1}$ *nếu* $x = hu$ *thì* $\mathfrak{I}(x) \subseteq \mathfrak{I}(u) \in I_k$;

(iv) Với $m \geq 1$, *tập* $\{\mathfrak{I}(y) | y = k_m ... k_1 x, \forall k_m, ..., k_1 \in H\}$ *gồm các khoảng mờ của các phần tử có độ dài* $m + |x|$ *là một phân hoạch trên* $\mathfrak{I}(x)$.

Những tính chất trên mô tả cấu trúc của tập các khoảng mờ, là cơ sở quan trọng để chúng ta đưa ra hàm định lượng ngữ nghĩa cho các giá trị ngôn ngữ. Hàm dấu $Sgn : \underline{X} \to \{-1, 0, 1\}$ (xem [12]) trong định nghĩa sau, xác định bởi tính âm dương (làm yếu đi hay mạnh thêm) của gia tử đối với phần tử sinh nguyên thủy cũng như đối với gia tử khác nhằm bảo toàn quan hệ thứ tự cho hàm định lượng ngữ nghĩa.

Định nghĩa 2.1. [12] Cho $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ là một ĐSGT tuyến tính đầy đủ, tự do; $fm(c^-)$, $fm(c^+)$ và $\mu(h)$ lần lượt là độ mờ của các phần tử sinh nguyên thủy c^-, c^+ và các gia tử $h \in H$, thỏa mãn các điều kiện 2) và 5) trong Mệnh đề 2.1. Khi đó, $v : \underline{X} \to [0, 1]$ được gọi là hàm định lượng ngữ nghĩa nếu nó được xác định như sau:

1) $v(W) = K = fm(c^-)$, $v(c^-) = K - \alpha fm(c^-) = \beta fm(c^-)$, $v(c^+) = K + \alpha fm(c^+)$;

2) $v(h_j x) = v(x) + Sgn(h_j x)\{\sum_{i=Sgn(j)}^{j} \mu(h_i) fm(x) - \omega(h_j x)\mu(h_j) fm(x)\}$, ở đây
$\omega(h_j x) = \frac{1}{2}[1 + Sgn(h_j x) Sgn(Sgn(h_p h_j x)(\beta - \alpha)] \in \{\alpha, \beta\}$, với mọi $j \in [-q^\wedge p]\}$;

3) $v(\Phi c^-) = 0$, $v(\Sigma c^-) = K = v(\Phi c^+)$, $v(\Sigma c^+) = 1$ và với mọi $j \in [-q^\wedge p]$, ta có:
$v(\Phi h_j x) = v(x) + Sgn(h_j x)\{\sum_{i=Sgn(j)}^{j-Sgn(j)} \frac{1+Sgn(h_j x)}{2} \mu(h_i) fm(x)\}$ và
$v(\Sigma h_j x) = v(x) + Sgn(h_j x)\{\sum_{i=Sgn(j)}^{j-Sgn(j)} \frac{1-Sgn(h_j x)}{2} \mu(h_i) fm(x)\}$.

Dựa trên cấu trúc của các khoảng mờ được mô tả trong Mệnh đề 2.2 và cách định lượng các giá trị ngôn ngữ trong Định nghĩa 2.1, chúng ta thấy có những điểm cần chú ý sau:

1) Giá trị $v(x)$ là điểm chia tập các phân hoạch $\{\mathfrak{I}_{k+1}(h_i x) | i \in [q^\wedge p]\}$ trên $\mathfrak{I}_k(x)$ thành hai phần gồm $\{\mathfrak{I}_{k+1}(h_i x) | i \in [1, p]\}$ và $\{\mathfrak{I}_{k+1}(h_i x) | i \in [-q, -1]\}$. Thành phần thứ nhất nằm bên phải điểm $v(x)$ nếu như $Sign(h_p x) = +1$ (tức $h_p x \geq x$), thành phần thứ hai nằm bên trái. Nếu $Sign(h_p x) = -1$ (tức $h_p x \leq x$) thì thành phần thứ nhất nằm bên trái, thành phần thứ hai nằm bên phải điểm $v(x)$. Điểm $v(x)$ cũng là điểm đầu mút chung của hai khoảng $\mathfrak{I}_{k+1}(h_1 x)$, $\mathfrak{I}_{k+1}(h_{-1} x)$; đồng thời nó cũng chia khoảng mờ $\mathfrak{I}_k(x)$ theo tỉ lệ $\alpha : \beta$ nếu $Sign(h_p x) = +1$ hoặc chia theo tỉ lệ $\beta : \alpha$ nếu $Sign(h_p x) = -1$. Từ điều này, ta suy ra $\sum_{1 \leq i \leq p} |\mathfrak{I}_{k+1}(h_i x)| = \beta |\mathfrak{I}_k(x)|$.

2) Trường hợp $|x| = j < k$, $v_A(x)$ luôn luôn là điểm đầu mút chung của hai khoảng $\mathfrak{I}_{k+1}(h_i y)$, $\mathfrak{I}_{k+1}(h_{i'} y')$ với chỉ số $i, i' \in \{-q, p\}$, tức là $v_A(x)$ nằm giữa hai khoảng này.

3. MÔ HÌNH QUAN HỆ CỦA CƠ SỞ DỮ LIỆU NGÔN NGỮ

3.1. Giới thiệu mô hình

Lược đồ cơ sở dữ liệu (CSDL) ngôn ngữ là một tập $DB = \{U, R_1, R_2, ..., R_m; Const\}$, ở đây $U = \{A_1, A_2, ..., An\}$ là miền trị của các thuộc tính, mỗi R_i là một lược đồ quan hệ, $Const$ là tập các ràng buộc dữ liệu. Gọi $\mathfrak{D}_{A_i}$ là miền trị của thuộc tính A_i. Trong CSDL ngôn ngữ, $\mathfrak{D}_{A_i} = D_{A_i} \cup LDom(A_i)$, trong đó D_{A_i} là miền trị tham chiếu và $LDom(A_i)$ là miền trị ngôn ngữ của thuộc tính A_i. Các giá trị trong $LDom(A_i)$ là các phần tử của ĐSGT tương ứng với A_i đóng vai trò như một biến ngôn ngữ có không gian cơ sở chính là D_{A_i}. Nếu $LDom(A_i) \neq \emptyset$ thì A_i được gọi là thuộc tính ngôn ngữ. Trong bài báo này, chúng ta sẽ xét CSDL ngôn ngữ mở rộng chứa các dạng dữ liệu khác nhau được được đề cập trong [17, 8, 9, 22]:

Kiểu 1: Giá trị ngôn ngữ mờ (tuổi rất trẻ)

Kiểu 2: Giá trị rõ (tuổi bằng 24 hoặc tên là Nam)

Kiểu 3: Giá trị khoảng (nhiệt độ trong ngày $25 \leq t \leq 35$)

Kiểu 4: Giá trị không xác định (undefine, inapplicable) (không có học vị khi xét thuộc tính học vị khoa học)

Kiểu 5: Dữ liệu thiếu vắng (missing) (lương chắc chắn có nhưng không biết bao nhiêu)

Kiểu 6: Tập hữu hạn các giá trị rõ (tuổi là một trong số các số thuộc $\{31, 33, 35\}$)

Kiểu 7: Giá trị "không biết" (unknown) (đã kết hôn nhưng không biết có con chưa, hay là đã có con nhưng không biết bao nhiêu con).

Trong CSDL thông thường, mỗi thông tin trên miền trị của các thuộc tính là thông tin chính xác nên nó được biểu diễn bởi một giá trị duy nhất trên không gian cơ sở. Tuy nhiên, điều này là không thể đối với các thông tin không chính xác trong CSDL ngôn ngữ mở rộng đã trình bày trên đây. Vì vậy, trước tiên chúng ta sẽ đưa ra quan điểm mới để biểu diễn cho dữ liệu kiểu 1, và trên cơ sở đó, sẽ biểu diễn các dạng dữ liệu còn lại một cách thống nhất.

3.2. Biểu diễn khoảng cho các giá trị ngôn ngữ trên miền trị của các thuộc tính

Biểu diễn giá trị ngôn ngữ mờ x bằng lát cắt α của tập mờ $A(x)$ thực chất là cách biểu diễn các khoảng rõ. Nếu $A(x)$ là tập mờ lồi, chuẩn tắc thì với mỗi $\alpha \in (0, 1]$, lát cắt $A^{\alpha}(x)$ là một khoảng trên không gian cơ sở. Hơn nữa, khi α dần về 1 thì các khoảng tương ứng bé dần và hội tụ về x. Vì thế, các khoảng này có thể được xem như các lân cận của x. Quan điểm này và khái niệm lân cận trên không gian tôpô dựa theo ngữ nghĩa trong ĐSGT có những điểm tương đồng. Đây chính là ý tưởng cho việc đề xuất cách biểu diễn mới cho các kiểu dữ liệu.

Như đã trình bày trong Mục 2, $H(x)$ là lân cận tôpô của x và họ $\mathfrak{V} = \{H(x)|x \in X\}$ là một cơ sở tôpô $\mathfrak{T}$ của X. Dễ thấy, họ các khoảng mờ $\mathfrak{B} = \{\mathfrak{I}(x)|x \in X\}$ đẳng cấu với cơ sở tôpô $\mathfrak{T}$. Vì vậy $\mathfrak{B}$ có thể xem là một cơ sở tôpô trên $[0, 1]$ - miền trị cơ sở đã được chuẩn hóa của thuộc tính. Bây giờ, giả sử x là một giá trị ngôn ngữ mờ của thuộc tính A, dựa trên cơ sở tôpô $\mathfrak{B}$ chúng ta sẽ xây dựng hệ lân cận của x.

Giá trị của hàm định lượng $v(x)$, tương tự như nhân của tập mờ, được xem là giá trị tương thích nhất cho ngữ nghĩa của x. Một lân cận của x là hợp các khoảng rõ trong $\mathfrak{B}$ mà mỗi khoảng này là một khoảng thuộc $\mathfrak{I}_{|x|}(x)$, nằm quanh $v(x)$. Nói cách khác, các lân cận của x là những khoảng trên không gian tham chiếu mà $v(x)$ là điểm trong theo nghĩa tôpô thông thường, tức là phải chứa một khoảng con đủ lớn các giá trị tương thích ngữ nghĩa với x ở một mức độ nào đó.

Xét ĐSGT đầy đủ, tuyến tính và tự do $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$, với $H^+ = \{h_1, ..., h_p\}$, $H^- = \{h_{-1}, ..., h_{-q}\}$. Giả thiết $h_1 < ... < h_p$ và $h_{-1} < ... < h_{-q}$, ở đây $p, q \geq 2$. Trong thực tế, số gia tử trong các giá trị ngôn ngữ là hữu hạn nên tồn tại một số nguyên dương k^* sao cho $0 < |x| \leq k^*$, $\forall x \in X$. Với bất kỳ $x \in X$, đặt $j = |x|$; với mỗi số nguyên k cho trước với $1 \leq k \leq k^*$, lân cận tối thiểu mức k của x được định nghĩa như sau:

a) Trường hợp $k = j$: khoảng mờ $\mathfrak{I}_k(x)$ là ảnh của $H(x)$ trên không gian tham chiếu. $H(x)$ đặc trưng cho tập tối đa các giá trị ngôn ngữ xấp xỉ với x về mặt ngữ nghĩa [11,12]. Nếu các giá trị trên miền trị tương ứng với tập này được coi là xấp xỉ nhau mức k thì sẽ có những cặp quá xa nhau và dường như không phù hợp với ý nghĩa của lân cận tối thiểu nên chúng ta sẽ sử dụng một số khoảng mờ mức $k+1$. Tập các $\mathfrak{I}_{k+1}(hx)$, $\forall h \in H$ là một phân hoạch của $\mathfrak{I}_k(x)$ và $v(x)$ là điểm đầu mút chung của các khoảng $\mathfrak{I}_{k+1}(h_{-1}x)$ và $\mathfrak{I}_{k+1}(h_1x)$. Vị trí của hai khoảng này có thể tráo đổi cho nhau tùy vào $h_{-1}x \leq x \leq h_1x$ hay là $h_1x \leq x \leq h_{-1}x$. Vì vậy, lân cận tối thiểu của x được định nghĩa là hợp của hai khoảng $\mathfrak{I}_{k+1}(h_{-1}x)$ và $\mathfrak{I}_{k+1}(h_1x)$:

$$O_{min,k}(x) = \mathfrak{I}_{k+1}(h_{-1}x) \cup \mathfrak{I}_{k+1}(h_1x) \tag{3.1}$$

b) Trường hợp $1 \leq k < j$: Giá trị của k chỉ cho mức mờ mà chúng ta đang quan tâm. Độ mờ của các phần tử trong X tỉ lệ nghịch với độ dài của nó. Khi $k < j$, mức mờ chúng ta xét để định nghĩa $O_{min,k}(x)$ lớn hơn mức mờ của bản thân dữ liệu x và khoảng mờ $\mathfrak{I}_j(x)$ chứa $v(x)$ bao giờ cũng được chứa trong một khoảng mờ mức k. Vì vậy, bản thân x có độ mờ nhỏ (vì $j > k$) nên nó không thể được mô tả bằng khoảng tính mờ mức k và do đó chúng ta sẽ phải lấy chính khoảng $\mathfrak{I}_j(x)$ làm lân cận tối thiểu mức k của x:

$$O_{min,k}(x) = \mathfrak{I}_j(x) \tag{3.2}$$

c) Trường hợp $j+1 \leq k \leq k^*$: Lập luận tương tự như trường hợp a) lân cận tối thiểu mức k của x được định nghĩa là hợp của hai khoảng mờ mức $k+1$, được xác định như sau. Theo chú ý cuối mục 2.2, $v(x)$ là đầu mút của các đoạn $\mathfrak{I}_{k+1}(h_l y)$ và $\mathfrak{I}_{k+1}(h_{l'} y')$ ở đây chọn $l, l' \in \{-q, p\}$ sao cho $h_l y \leq x \leq h_{l'} y'$; và $y, y' \in H(x)$ với $|y| = |y'| = k$. Như vậy,

$$O_{min,k}(x) = \mathfrak{I}_{k+1}(h_l y) \cup \mathfrak{I}_{k+1}(h_{l'} y') \tag{3.3}$$

Cuối cùng, chúng ta thống nhất cách biểu diễn dữ liệu ngôn ngữ mờ theo định nghĩa sau đây.

Định nghĩa 3.1. Cho $x \in X \cup C$, một *biểu diễn khoảng* của x là một tập $IRp(x)$ các khoảng được xác định như sau:

$$IRp(x) = \{O_{min,k}(x) | 1 \leq k \leq k^*\} \tag{3.4}$$

Ví dụ 3.1 Cho ĐSGT tuyến tính của biến Tuổi là $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \leq)$ với $G = \{trẻ, già\}$, $H^- = \{gần, ít\}$ và $H^+ = \{khá, rất\}$. Chú ý là $rất > khá$ và $ít > gần$. Giả sử miền trị tham chiếu của biến Tuổi của những người đang công tác là $D_A = [18, 60]$; các tham số mờ $fm(già) = 0.58$, $fm(trẻ) = 0.42$, $\mu(gần) = 0.27$, $\mu(ít) = 0.25$, $\mu(khá) = 0.28$, $\mu(rất) = 0.20$; suy ra $\alpha = 0.52$, $\beta = 0.48$. Chúng ta sẽ biểu diễn giá trị ngôn ngữ $x = khá\text{-}trẻ$.

Giả sử $k^* = 3$. Ta có $|x| = 2$ và $D_A = [18, 60]$ nên sẽ dùng hệ số r để chuyển đổi từ [0,1] qua [18, 60], các kí hiệu có kèm r chỉ cho sự chuyển đổi này.

1) Trường hợp $k < 2 = |khá\text{-}trẻ|$: Ta có $|\mathfrak{I}_{1,r}(trẻ)| = fm(trẻ) \times (60 - 18) = 0.42 \times 42 = 17.64$. Suy ra $\mathfrak{I}_{1,r}(trẻ) = [18.00, 35.64]$ và $fm_r(khá\text{-}trẻ) = \mu(khá) \times fm(trẻ) \times (60-18) = 0.28 \times 0.42 \times 42 = 4.9392$; $fm_r(rất\text{-}trẻ) = \mu(rất) \times fm(trẻ) \times (60 - 18) = 0.20 \times 0.42 \times 42 = 3.528$.

Vì $\{\mathfrak{I}_{2,r}(rất\text{-}trẻ), \mathfrak{I}_{2,r}(khá\text{-}trẻ), \mathfrak{I}_{2,r}(gần\text{-}trẻ), \mathfrak{I}_{2,r}(ít\text{-}trẻ)\}$ là một phân hoạch của $\mathfrak{I}_{1,r}(trẻ)$, ta suy ra $\mathfrak{I}_{2,r}(rất\text{-}trẻ) = [18, 21.5280]$ và $\mathfrak{I}_{2,r}(khá\text{-}trẻ) = (21.5280, 26.4672]$. Vậy, $O_{min,1}(khá\text{-}trẻ) = \mathfrak{I}_{2,r}(khá\text{-}trẻ) = (21.5280, 26.4672]$.

2) Trường hợp $k = j = 2$: Ta có $H(khá\text{-}trẻ) \cap X_3 = \{rất\text{-}khá\text{-}trẻ, khá\text{-}khá\text{-}trẻ, gần\text{-}khá\text{-}trẻ, ít\text{-}khá\text{-}trẻ\}$, có thứ tự từ nhỏ đến lớn như đã liệt kê. Ta có $O_{min,2}(khá\text{-}trẻ) = \mathfrak{I}_{3,r}(khá\text{-}khá\text{-}trẻ) \cup \mathfrak{I}_{3,r}((gần\text{-}khá\text{-}trẻ)$. Các khoảng mờ liên quan được tính như sau: $fm_r(rất\text{-}khá\text{-}trẻ) = \mu(rất) \times fm_r(khá\text{-}trẻ) = 0.20 \times 4.9392 = 0.98784$. Vì vậy, $\mathfrak{I}_{3,r}(rất\text{-}khá\text{-}trẻ) = (21.5280, 22.51584]$. Tương tự, ta có $\mathfrak{I}_{3,r}(khá\text{-}khá\text{-}trẻ) = (22.51584, 23.89882]$ và $\mathfrak{I}_{3,r}(gần\text{-}khá\text{-}trẻ) = (23.89882, 25.2324]$. Do đó $O_{min,2}(khá\text{-}trẻ) = (22.51584, 25.2324]$.

3) Trường hợp $k = 3$: $O_{min,3}(khá\text{-}trẻ) = \mathfrak{I}_{4,r}(ít\text{-}khá\text{-}khá\text{-}trẻ) \cup \mathfrak{I}_{4,r}(ít\text{-}gần\text{-}khá\text{-}trẻ)$. Vì $v(khá\text{-}trẻ) = 23.89882$ và $fm_r(ít\text{-}khá\text{-}khá\text{-}trẻ) = \mu(ít) \times fm_r(khá\text{-}khá\text{-}trẻ) = 0.25 \times 1.382976 = 0.345744$; $fm_r(rất\text{-}gần\text{-}khá\text{-}trẻ) = \mu(rất) \times fm_r(gần\text{-}khá\text{-}trẻ) = 0.2 \times 1.333584 = 0.333396$ nên $\mathfrak{I}_{4,r}(ít\text{-}khá\text{-}khá\text{-}trẻ) = (23.55308, 23.89882]$, $\mathfrak{I}_{4,r}(ít\text{-}gần\text{-}khá\text{-}trẻ) = (23.89882, 24.23222]$ và do đó $O_{min,3}(khá\text{-}trẻ) = (23.55308, 24.23222]$.

Như vậy, $IRp(khá\text{-}trẻ) = \{(21.5280, 26.4672], (22.51584, 25.2324], (23.55308, 24.23222]\}$.

3.3. Biểu diễn khoảng cho các dạng dữ liệu khác

Cách biểu diễn dữ liệu ngôn ngữ mờ trên đây có thể sử dụng để biểu diễn các dạng dữ liệu còn lại đã nói trong Mục 3.1:

a) Kiểu 2: Mỗi giá trị thực a là dữ liệu rõ, độ mờ của dữ liệu bằng 0, sẽ được biểu diễn bằng $[a, a]$, tương ứng với mức mờ luôn luôn là ∞ nên còn gọi là khoảng mờ mức ∞ của a. Vì vậy, $O_{min,k}(a) = \{[a, a]\}$, với mọi $1 \leq k \leq k^*$ và $IRp(a) = \{[a, a]\}$.

b) Kiểu 3: Mỗi giá trị khoảng $[a, b]$ được biểu diễn bằng một tập chứa duy nhất khoảng $[a, b]$. Vì $[a, b]$ là dữ liệu rõ nên $O_{min,k}([a, b]) = \{[a, b]\}$, với mọi $1 \leq k \leq k^*$ và $IRp([a, b]) = \{[a, b]\}$.

c) Kiểu 4: Mỗi giá trị không được xác định (undefine, inapplicable) được biểu diễn bằng tập $\emptyset$, xem như thông tin chính xác. Vì vậy $O_{min,k}(inapplicable) = \{\emptyset\}$, với mọi $1 \leq k \leq k^*$ và $IRp(inapplicable) = \{\emptyset\}$.

d) Kiểu 5: Giá trị kiểu này có nhưng không biết là bao nhiêu (missing) nghĩa là có thể nhận bất kỳ một giá trị rõ nào đó trên miền trị của thuộc tính. Theo quan điểm đó,

$O_{min,k}(missing) = \{[a,a]|a \in D\}$, với mọi $1 \le k \le k^*$ và $IRp(missing) = \{[a,a]|a \in D_A\}$.

e) Kiểu 6: Giá trị kiểu này có thể là một giá trị thuộc một tập $P \subseteq D_A$ nhưng chưa biết là giá trị nào. Tương tự như kiểu 5, $O_{min,k}(P) = \{[a,a]|a \in P\}$, với mọi $1 \le k \le k^*$ và $IRp(P) = \{[a,a]|a \in P\}$.

f) Kiểu 7: Về mặt ý nghĩa, giá trị kiểu này (unknown) có thể xem là sự kết hợp của kiểu 4 và kiểu 5. Do vậy $IRp(unknown) = \{\emptyset, [a,a]|a \in D_A\}$.

Với phương pháp biểu diễn khoảng vừa nêu trên, chúng ta đã xem các kiểu dữ liệu khác nhau trên một quan điểm thống nhất. Mỗi thông tin đều được biểu diễn bởi một tập các khoảng trên không gian tham chiếu. Các tham số cần xác định bao gồm: độ mờ của phần tử sinh nguyên thủy, của các gia tử và giá trị k^*. So với cách biểu diễn bằng tập mờ hoặc phân phối khả năng thì số lượng các tham số ít hơn nhiều và người dùng có trực quan để có thể xác định dễ dàng hơn so với cách xác định tập mờ cho mọi giá trị ngôn ngữ.

4. MỘT CƠ SỞ HÌNH THỨC THỐNG NHẤT CHO CÁC THAO TÁC TRÊN NGỮ NGHĨA CỦA DỮ LIỆU

Một vấn đề được đặt ra khi xử lý các CSDL ngôn ngữ là định nghĩa quan hệ bằng nhau như thế nào để so sánh hai thông tin không chính xác. Đã có nhiều nghiên cứu về quan hệ này. Chúng được gọi là quan hệ tương tự [4-6], quan hệ xấp xỉ, quan hệ gần nhau [7,18,19]. Từ đó, người ta định nghĩa các quan hệ đối sánh khác $<$, $>$, $\le$ và $\ge$ trên miền trị mở rộng của các thuộc tính. Trong mục này, chúng ta sẽ trình bày một cơ sở toán học để định nghĩa các quan hệ như trên với mục đích cuối cùng là có thể xử lý các truy vấn dữ liệu dễ dàng, theo một cách thống nhất.

4.1. Quan hệ gần nhau mức k và bằng nhau mức k trên miền trị của thuộc tính

Quan hệ gần nhau sẽ được xây dựng dựa trên các khoảng mờ của các phần tử trong X. Chúng là một cơ sở tôpô ngữ nghĩa trên miền trị D_A như đã trình bày trong Mục 2. Vì tập các khoảng mờ thuộc I_k là một phân hoạch trên D_A nên nó xác định một quan hệ tương đương với các lớp tương đương là các khoảng mờ này. Các giá trị nằm trong cùng khoảng sẽ được coi là gần nhau mức k. Tuy nhiên, như đã nhận xét trong Mục 2.2, khi x có độ dài bé hơn k thì giá trị $v(x)$ là điểm đầu mút của một lớp tương đương $\mathfrak{I}_k(u)$ trong I_k. Điều này dẫn đến có những giá trị trong lân cận của x lại không tương tự mức k. Vì vậy, chúng ta sẽ xây dựng một phân hoạch khác sao cho $v(x)$ là điểm trong tôpô của phân hoạch với mọi x, $|x| \le k$, như sau:

Xét ĐSGT tuyến tính, đầy đủ và tự do $\underline{AX} = (\underline{X}, G, C, H, \Phi, \Sigma, \le)$ với $H^+ = \{h_1, ..., h_p\}$, $H^- = \{h_{-1}, ..., h_{-q}\}$ với $h_1 < ... < h_p$ và $h_{-1} < ... < h_{-q}$, ở đây $p, q \ge 2$. Đặt H_1 là tập các gia tử yếu, H_2 là tập các gia tử mạnh theo nghĩa khi tác động nó sẽ làm thay đổi nghĩa mạnh hơn số gia tử trong H_1, tức là các tập H_1 và H_2 gồm:

$$H_1 = \{h_i, h_{-j}|1 \le i \le [p/2], 1 \le j \le [q/2]\}, \tag{4.1a}$$

$$H_2 = \{h_i, h_{-j}|[p/2] < i \le p, [q/2] < j \le q\}. \tag{4.1b}$$

Đặt $I_{k+1}(H_n) = \{\Im_{k+1}(h_i y) | y \in X_k, h_i \in H_n\}$, với $n = 1, 2$. Hai khoảng $\Im_{k+1}(x)$ và $\Im_{k+1}(y)$ trong $I_{k+1}(H_n)$ được gọi là liên thông nhau với nhau nếu tồn tại các khoảng thuộc $I_{k+1}(H_n)$ liên tiếp nhau xếp từ $\Im_{k+1}(x)$ đến $\Im_{k+1}(y)$. Quan hệ này sẽ phân $I_{k+1}(H_n)$ thành các thành phần liên thông. Theo (2.1) và (2.2), với mỗi $y \in X_k$, ta thấy $I_{k+1}(H_1)$ được phân thành các cụm (cluster) có dạng $\{\Im_{k+1}(h_i y) | h_i \in H_1\}$. Hơn nữa, do $\Im_{k+1}(h_{-1} y) \leq v(y) \leq \Im_{k+1}(h_1 y)$ hoặc là $\Im_{k+1}(h_1 y) \leq v(y) \leq \Im_{k+1}(h_{-1} y)$ nên bao giờ ta cũng có $v(y) \in \{\Im_{k+1}(h_i y) | h_i \in H_1\}$.

Bây giờ ta phân cụm các khoảng mờ của $I_{k+1}(H_2)$. Giả sử $X_k = \{x_s | s = 0, ..., m-1\}$ gồm m phần tử được sắp thành một dãy sao cho $x_i \leq x_j$ khi và chỉ khi $i \leq j$. Kí hiệu $H_2^- = H_2 \cap H^-$ và $H_2^+ = H_2 \cap H^+$. Để ý rằng $h_{-q} \in H_2^-$ và $h_p \in H_2^+$. Các cụm được sinh ra từ các khoảng mờ thuộc $I_{k+1}(H_2)$ có ba loại sau đây:

1) Cụm nằm bên trái x_0: Các khoảng mờ của $\Im_{k+1}(h_i x_0)$ với $i \in [-q \wedge p]$ sắp xếp trên khoảng $\Im_k(x_0)$ theo thứ tự trong (2.2), do $Sgn(h_p x_0) = -1$. Vì vậy, cụm bên trái của x_0 là $\{\Im_{k+1}(h_i x_0) | h_i \in H_2^+\}$.

2) Cụm nằm bên phải s_{m-1}: Các khoảng mờ của $\Im_{k+1}(h_i x_{m-1})$ với $i \in [-q \wedge p]$ sắp xếp trên khoảng $\Im_k(x_{m-1})$ theo thứ tự trong (2.1), do $Sgn(h_p x_{m-1}) = +1$. Vì vậy, cụm bên phải của x_{m-1} là $\{\Im_{k+1}(h_i x_{m-1}) | h_i \in H_2^+\}$.

3) Các cụm nằm giữa x_s và x_{s+1} với $s = 0, ..., m-2$: phụ thuộc vào $Sgn(h_p x_s)$ và $Sgn(h_p x_{s+1})$ như sau

$$\mathcal{C} = \{\Im_{k+1}(h_i x_s), \Im_{k+1}(h'_j x_{s+1}) | h_i \in H_2^+, h'_j \in H_2^-\}, \tag{4.2a}$$

nếu $Sgn(h_p x_s) = +1$ và $Sgn(h_p x_{s+1}) = +1$.

$$\mathcal{C} = \{\Im_{k+1}(h_i x_s), \Im_{k+1}(h'_j x_{s+1}) | h_i \in H_2^+, h'_j \in H_2^+\}, \tag{4.2b}$$

nếu $Sgn(h_p x_s) = +1$ và $Sgn(h_p x_{s+1}) = -1$.

$$\mathcal{C} = \{\Im_{k+1}(h_i x_s), \Im_{k+1}(h'_j x_{s+1}) | h_i \in H_2^-, h'_j \in H_2^-\}, \tag{4.2c}$$

nếu $Sgn(h_p x_s) = -1$ và $Sgn(h_p x_{s+1}) = +1$.

$$\mathcal{C} = \{\Im_{k+1}(h_i x_s), \Im_{k+1}(h'_j x_{s+1}) | h_i \in H_2^-, h'_j \in H_2^+\}, \tag{4.2d}$$

nếu $Sgn(h_p x_s) = -1$ và $Sgn(h_p x_{s+1}) = -1$.

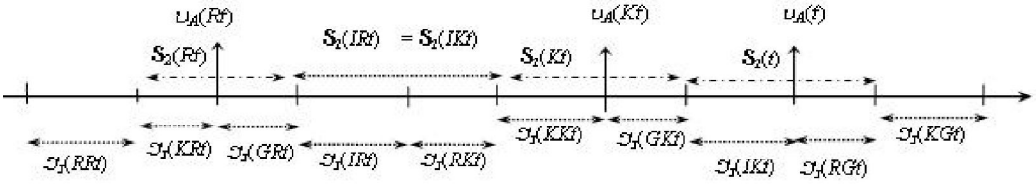
Tập tất cả các cụm được ký hiệu là $\mathfrak{C}$ và ta sẽ định nghĩa khoảng tương tự mức k như sau.

Định nghĩa 4.1. Mỗi $\mathcal{C}$ thuộc $\mathfrak{C}$, ta gọi khoảng tương tự mức k ứng với $\mathcal{C}$ là

$$\mathcal{S}_k(\mathcal{C}) = \bigcup \{\Im_{k+1} | \Im_{k+1} \in \mathcal{C}\} \tag{4.3}$$

Với cách định nghĩa này, mỗi khoảng $\mathcal{S}_k(\mathcal{C})$ sẽ không quá lớn để phủ bất kỳ một khoảng $\Im_k$ nhưng lại cũng không quá nhỏ để nằm gọn trong một khoảng $\Im_{k+1}$ nào. Vì $\{\mathcal{S}_k(\mathcal{C}))|\mathcal{C} \in \mathfrak{C}\}$

là một phân hoạch trên miền trị tham chiếu nên nó xác định một quan hệ tương đương và chúng ta sẽ gọi là quan hệ tương tự mức k. Do tính chất của phân hoạch nên với mỗi giá trị x của thuộc tính, tồn tại duy nhất một cụm $\mathcal{C}$ sao cho $v(x) \in \mathcal{S}_k(\mathcal{C})$. Vì vậy, chúng ta có thể định nghĩa $\mathcal{S}_k(x) = \mathcal{S}_k(\mathcal{C})$.



Hình 4.1. Phân cụm một số khoảng mờ của biến ngôn ngữ "Tuổi"

Để minh họa cho khái niệm này, chúng ta xét một số khoảng mờ trong Hình 4.1. Trong đó các gia tử "It", "Gần", "Khá", "Rất" được kí hiệu là I, G, K, R và "trẻ" được kí hiệu là t. Với $k = 2$, ta có $\{\mathfrak{I}_3(KRt), \mathfrak{I}_3(GRt)\}$ và $\{\mathfrak{I}_3(KKt), \mathfrak{I}_3(GKt)\}$ là các cụm thuộc $I_3(H_1)$; $\{\mathfrak{I}_3(RRt)\}$, $\{\mathfrak{I}_3(IRt), \mathfrak{I}_3(RKt)\}$ và $\{\mathfrak{I}_3(IKt), \mathfrak{I}_3(RGt)\}$ là các cụm thuộc $I_3(H_2)$. Từ đó suy ra các khoảng $\mathcal{S}_2$ như trong Hình 4.1.

Mệnh đề 4.1. *Cho $\underline{AX}$ là ĐSGT tuyến tính đầy đủ của thuộc tính A, trong đó H^+ và H^- có ít nhất hai phần tử, các tham số định lượng mờ được xác định theo Định nghĩa 2.1. Khi đó*:

a) *Với mỗi k, $\{\mathcal{S}_k(u) | u \in X \cup C\}$ được xác định duy nhất và là một phân hoạch của đoạn* $[0, 1]$.

b) *Với mọi $x, u \in X \cup C$, nếu $v(x) \in \mathcal{S}_k(u)$ thì lân cận bé nhất mức k của x nằm trong $\mathcal{S}_k(u)$, tức là $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$.*

Chứng minh. a) Tập các khoảng tính mờ thuộc I_{k+1} là một phân hoạch trên $[0, 1]$ và các cụm trong $\mathfrak{C}$ là một phân hoạch của I_{k+1}. Hơn nữa, I_{k+1} xác định duy nhất bởi các tham số mờ. Từ đó ta suy ra khẳng định (a).

b) Giả sử $v(x) \in \mathcal{S}_k(u)$, $|x| = j$ và $x = h_{j-1}...h_1c$ là biểu diễn chính tắc của x đối với $c \in \{c^-, c^+\}$.

Trường hợp $k \leq j$: Nếu $j = 1$ thì $x \in \{0, c^-, W, c^+, 1\}$ và $k = j = 1$. Giả sử $x = c \in \{c^-, c^+\}$, từ $v(x) \in \mathcal{S}_1(u)$ ta suy ra $\mathcal{S}_1(u)$ là hợp của các khoảng sinh bởi các gia tử trong H_1, tức là $\mathcal{S}_1(u) = S_1(\mathcal{C}) = \cup\{\mathfrak{I}_2(h_i'c) | h_i' \in H_1\}$. Vì $O_{min,1}(x) = \mathfrak{I}_{k+1}(h_{-1}c) \cup \mathfrak{I}_{k+1}(h_1c)$ nên $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$. Đối với $x \in \{0, W, 1\}$, chứng minh tương tự.

Nếu $1 \leq k < j$ thì $O_{min,k}(x) = \mathfrak{I}_j(x)$. Ký hiệu $x_{|k} = h_{k-1}...h_1c$. Nếu $h_k \in H_1$ thì từ $v(x) \in \mathcal{S}_k(u)$ ta suy ra $\mathcal{S}_k(u) = \cup\{\mathfrak{I}_{k+1}(h_i'x_{|k}) | h_i' \in H_1\}$. Theo Mệnh đề 2.2(iv), $\mathfrak{I}_j(x) \subseteq \mathfrak{I}_{k+1}(h_kx_{|k}) \subseteq \mathcal{S}_k(u)$. Nếu $h_k \in H_2^-$ thì theo (4.3a)-(4.3d) và Định nghĩa của $\mathcal{S}_k$, tồn tại y có độ dài k sao cho $\mathcal{S}_k(u) = \cup\{\mathfrak{I}_{k+1}(h_i'x_{|k}), \mathfrak{I}_{k+1}(h_j'y) | h_i' \in H_2^-, h_j' \in H_2^\epsilon\}$, ở đây $\epsilon \in \{-, +\}$. Dễ thấy $\mathfrak{I}_j(x) \subseteq \mathfrak{I}_{k+1}(h_kx_{|k}) \subseteq \mathcal{S}_k(u)$, tức là $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$. Trường hợp $h_k \in H_2^+$, chứng minh tương tự.

Trường hợp $k = j$: Ta có $O_{min,k}(x) = \Im_{j+1}(h_{-1}x) \cup \Im_{j+1}(h_1x)$. Vì $v(x) \in \mathcal{S}_k(u)$ nên theo định nghĩa, $\mathcal{S}_k(u) = \cup\{\Im_{k+1}(h'_ix)|h'_i \in H_1\}$. Do đó $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$.

Trường hợp $k > j$: Theo định nghĩa, ta có $O_{min,k}(x) = \Im_{k+1}(h_ly) \cup \Im_{k+1}(h_{l'}y')$, ở đây $l, l' \in \{p, -q\}$, $y, y' \in H(x) \cap X_k$ và x nằm giữa h_ly và $h_{l'}y'$, tức là $h_ly \leq x \leq h_{l'}y'$ hoặc $h_ly \geq x \geq h_{l'}y'$. Giả sử $h_l \in H_2^{\epsilon}, h_{l'} \in H_2^{\epsilon'}$ với $\epsilon, \epsilon' \in \{-, +\}$, từ giả thiết của $v(x)$ ta suy ra $\mathcal{S}_k(u) = \cup\{\Im_{k+1}(h_iy), \Im_{k+1}(h'_iy')|h_i \in H_2^{\epsilon}, h'_i \in H_2^{\epsilon'}\}$. Vì vậy $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$. ■

Bây giờ chúng ta có thể định nghĩa khái niệm bằng nhau mờ mức k tổng quát như sau.

Định nghĩa 4.2. Cho ĐSGT tuyến tính, đầy đủ $\underline{AX}$ và độ đo mờ fm. Giả sử v_A là một hàm định lượng ngữ nghĩa trên $\underline{AX}$ và; với mỗi k mà $1 \leq k \leq k^*$, $\mathcal{S}_k$ là quan hệ tương tự mức k trên D_A. Khi đó, với hai bộ t và s tùy ý trên U, hai giá trị $t[A]$ và $s[A]$ trên miền trị $\mathcal{D}_A$ được gọi là bằng nhau mức k, kí hiệu là $t[A] =_{fm,k} s[A]$ hoặc $t[A] =_k s[A]$, nếu tồn tại một lớp tương đương $\mathcal{S}_k(u)$ của $\mathcal{S}_k$ sao cho $O_{min,k}(t[A]) \subseteq \mathcal{S}_k(u)$ và $O_{min,k}(s[A]) \subseteq \mathcal{S}_k(u)$.

Để minh họa, trong Hình 4.1 chúng ta thấy $KRt =_2 GRt =_2 Rt$ và $IKt =_2 RGt$.

4.2. Quan hệ bằng nhau mức k và quan hệ đối sánh trên miền trị mở rộng của thuộc tính

Như đã trình bày trong Mục 3.3, miền trị mở rộng của thuộc tính A sẽ bao gồm thêm các giá trị có kiểu từ 3 đến 7 và được ký hiệu là $\underline{\mathcal{D}}_A$. Bởi vì mỗi dữ liệu d thuộc các kiểu này đều có $O_{min,k}(d) = IRp(d)$ với mọi $0 < k \leq k^*$ nên $O_{min,k}(d) \subseteq \mathcal{S}_k(u)$ được định nghĩa như sau:

$$O_{min,k}(d) \subseteq \mathcal{S}_k(u) \text{ khi và chỉ khi } \forall \Im \in O_{min,k}(d) \text{ và } \Im \neq \emptyset, \text{ ta có } \Im \subseteq \mathcal{S}_k(u) \tag{4.4}$$

Theo Mệnh đề 4.1, giá trị ngôn ngữ x có $v(x) \in \mathcal{S}_k(u)$ sẽ tương tự mức k đối với các giá trị khác trong $\mathcal{S}_k(u)$. Điều kiện $\Im \neq \emptyset$ được đặt ra để phù hợp với ý nghĩa trực quan là dữ liệu "inapplicable", có biểu diễn $\{\emptyset\}$, không thể tương tự mức k với dữ liệu trong $\mathcal{S}_k(u)$. Từ Định nghĩa 4.2 và khái niệm bao hàm trong (4.4), ta đưa ra định nghĩa sau.

Định nghĩa 4.3. Với giả thiết như trong Định nghĩa 4.2, hai giá trị $t[A], s[A]$ thuộc $\underline{\mathcal{D}}_A$ được gọi là bằng nhau mức k, kí hiệu $t[A] =_k s[A]$, nếu một trong các điều kiện sau đây thỏa mãn:

(a) $t[A]$ và $s[A]$ đồng nhất nhau về kí hiệu;

(b) Tồn tại một lớp tương đương $\mathcal{S}_k(u)$ của quan hệ tương tự $\mathcal{S}_k$ sao cho $O_{min,k}(t[A]) \subseteq \mathcal{S}_k(u)$ và $O_{min,k}(s[A]) \subseteq \mathcal{S}_k(u)$.

Dữ liệu d được gọi là **tương thích được mức** k nếu như lân cận tối thiểu mức k của d nằm trọn trong một lớp tương đương của quan hệ $\mathcal{S}_k$. Theo định nghĩa này, mỗi phần tử $d \in D_A$ tương thích được mức k với k tùy ý, vì lân cận tối thiểu mức k của d luôn là $[d, d]$. Theo Mệnh đề 4.1, ta thấy với k tùy ý cố định trước, mọi giá trị ngôn ngữ đều tương thích được mức k. Tuy nhiên, các giá trị kiểu 3, 4, 7 không tương thích được mức k.

Nếu một trong các $t[A], s[A]$ không tương thích được mức k hoặc chúng đều tương thích nhưng lân cận tối thiểu mức k không nằm chung trong một lớp tương đương của $=_k$ thì ta nói $t[A] \neq_k s[A]$. Theo Định nghĩa này, với bất kỳ $d \in \underline{\mathcal{D}}_A$, ta có thể kiểm tra được

$d \neq_k$ *inapplicable*, $d \neq_k$ *missing*, $d \neq_k$ *unknown* ngoại trừ trường hợp chúng thỏa điều kiện (a). Trường hợp dữ liệu đoạn $[a, b]$ kiểu 3 có độ dài đủ lớn, ta cũng sẽ có $[a, b] \neq_k d$, $\forall d \in LDom(A)$. Kết quả cũng như vậy đối với dữ liệu kiểu 6. Từ Mệnh đề 4.1 và Định nghĩa 4.2, ta có:

Hệ quả 4.1. *Quan hệ* $=_k$ *là quan hệ tương đương trên miền trị mở rộng* $\underline{\mathcal{D}}_A$ *và với hai bộ* s, t *tùy ý thuộc* U *hoặc là* $t[A] =_k s[A]$ *hoặc là* $t[A] \neq_k s[A]$.

Bây giờ, chúng ta sẽ định nghĩa các quan hệ đối sánh còn lại trên $\underline{\mathcal{D}}_A$ như sau.

Định nghĩa 4.4. Với mỗi $k, 1 \leq k \leq k^*$ và hai giá trị bất kỳ $t[A], s[A]$ trong $\underline{\mathcal{D}}_A$, ta viết:

(*a*) $t[A] \leq_k s[A]$ khi và chỉ khi $t[A] =_k s[A]$ hoặc là $t[A], s[A]$ đều tương thích được mức k và $\mathcal{S}_k(t[A]) \leq \mathcal{S}_k(s[A])$;

(*b*) $t[A] <_k s[A]$ khi và chỉ khi $t[A], s[A]$ đều tương thích được mức k và $\mathcal{S}_k(t[A]) < \mathcal{S}_k(s[A])$.

Mệnh đề 4.2.

a) *Với mọi* $x \in X \cup C$ *và* k *tùy ý thỏa mãn* $1 \leq k \leq k^*$, *ta có* x *luôn tương thích được mức* k;

b) *Với hai giá trị tùy ý* $t[A], s[A] \in \underline{\mathcal{D}}_A$, *nếu chúng đều tương thích được mức* k *thì chỉ xảy ra một trong ba quan hệ sau đây: 1)* $t[A] =_k s[A]$; *2)* $t[A] <_k s[A]$; *3)* $t[A] >_k s[A]$.

Chứng minh.

a) Vì các lớp tương đương của $\mathcal{S}_k$ là một phân hoạch trên D_A nên tồn tại $u \in X \cup C$ sao cho $v(x) \in \mathcal{S}_k(u)$. Theo Mệnh đề 4.1(b), ta có $O_{min,k}(x) \subseteq \mathcal{S}_k(u)$, tức là x tương thích được mức k.

b) Giả sử $t[A], s[A] \in \underline{\mathcal{D}}_A$ đều tương thích được mức k, tức là $O_{min,k}(t[A]) \subseteq \mathcal{S}_k(u_t)$ và $O_{min,k}(s[A]) \subseteq \mathcal{S}_k(u_s)$, với $u_t, u_s \in X \cup C$. Theo Định nghĩa 4.3(b), Định nghĩa 4.4 và do các lớp tương đương của $\mathcal{S}_k$ là một phân hoạch trên D_A nên ta suy ra chỉ xảy ra một trong ba quan hệ trên. ■

Với mỗi $k, 1 \leq k \leq k^*$, đặt $X_{(k)} = X_1 \cup ... \cup X_k$, là tập tất cả các phần tử có độ dài tối đa bằng k. Về quan hệ giữa các $\mathcal{S}_k(x)$ với $x \in X_{(k)}$ ta có một số tính chất sau.

Mệnh đề 4.3. *Giả sử* $\mathcal{S}_k$ *là một quan hệ tương tự mức* k *trên* $LDom(A)$. *Ta có các khẳng định sau đây:*

a) *Với mọi* $x, y \in X_k$ *và* $x \neq y$ *ta có* $v(x) \notin \mathcal{S}_k(y)$ *và* $v(y) \notin \mathcal{S}_k(x)$.

b) *Với mọi* $x \in X_k$, *với mọi* $y \in X_{(k-1)} \cup C$ *ta có* $v(x) \notin \mathcal{S}_k(y)$ *và* $v(y) \notin \mathcal{S}_k(x)$.

Chứng minh

a) Vì $\mathcal{S}_k(x) = \{\mathfrak{I}_{k+1}(hx) | h \in H_1\}$ và $\mathfrak{I}_{k+1}(hx) \subseteq \mathfrak{I}_k(x)$ theo (iii) của Mệnh đề 2.1 nên $\mathcal{S}_k(x) \subseteq \mathfrak{I}_k(x)$. Hơn nữa, theo (ii) của Mệnh đề 2.1, $\{\mathfrak{I}_k(x) | x \in X_k\}$ là một phân hoạch trên $[0, 1]$ nên với mọi $x, y \in X_k$ ta có $\mathcal{S}_k(x) \cap \mathcal{S}_k(y) = \emptyset$. Vì $v(x) \in \mathcal{S}_k(x)$ và $v(y) \in \mathcal{S}_k(y)$ nên $v(x) \notin \mathcal{S}_k(y)$ và $v(y) \notin \mathcal{S}_k(x)$.

b) Giả sử $X_k = \{x_0, ..., x_m\}$ trong đó $x_0 < ... < x_m$. Với mỗi $y \in X_{(k-1)}$ hoặc $y = W$ tồn tại $s \in \{0, ..., m-1\}$ sao cho $x_s < y < x_{s+1}$. Theo (4.2a)-(4.2d), $\mathcal{S}_k(y) = \{\mathfrak{I}_{k+1}(h_l x_s) | h_l \in$

$H_2, h_l x_s > x_s\} \cup \{\Im_{k+1}(h_{l'} x_{s+1}) | h_{l'} \in H_2, h_{l'} x_{s+1} < x_{s+1}\}$. Với bất kỳ $x \in X_k$ ta có $\mathcal{S}_k(x) = \{\Im_{k+1}(hx) | h \in H_1\}$. Do $H_1 \cap H_2 = \emptyset$ nên hx, $h_l x_s$ cũng như hx, $h_{l'} x_{s+1}$ trong các biểu diễn trên đôi một khác nhau. Hơn nữa, vì $\{\Im_{k+1}(z) | z \in X_{k+1}\}$ là một phân hoạch trên $[0, 1]$ theo (ii) của Mệnh đề 2.2 nên ta suy ra $\mathcal{S}_k(x) \cap \mathcal{S}_k(y) = \emptyset$. Vậy $v(y) \notin \mathcal{S}_k(x)$ và $v(x) \notin \mathcal{S}_k(y)$.

Nếu $y = 0$ thì $\mathcal{S}_k(0) = \{\Im_{k+1}(h_l x_0) | h_l \in H_2, h_l x_0 < x_0\}$. Lập luận tương tự như trên ta được $v(0) \notin \mathcal{S}_k(x)$ và $v(x) \notin \mathcal{S}_k(0)$ với mọi $x \in X_k$. Trường hợp $y = 1$ chứng minh tương tự. ■

Hệ quả 4.2. *Mỗi lớp $\mathcal{S}_k(\mathcal{C})$ của quan hệ tương tự $\mathcal{S}_k$ có không quá một giá trị $x \in X_{(k)}$ sao cho $v(x) \in \mathcal{S}_k(\mathcal{C})$.*

Chứng minh. Chúng ta sẽ chứng minh bằng quy nạp theo mức k.

Bước cơ sở $k = 1$: Xét các phần tử trong $X_1 = \{c^-, c^+\}$ và $C = \{0, W, 1\}$. Theo Mệnh đề 4.3(a), $\mathcal{S}_1(c^-) \cap \mathcal{S}_1(c^+) = \emptyset$. Theo Mệnh đề 4.3(b), $v(c_0) \notin \mathcal{S}_1(c^-) \cup \mathcal{S}_1(c^+)$ với mọi $c_0 \in C$. Hơn nữa, $0 < c^- < W < c^+ < 1$ nên mỗi khoảng $\mathcal{S}_k(\mathcal{C})$ có duy nhất một giá trị $x \in X_1 \cup C$ sao cho $v(x) \in \mathcal{S}_k(\mathcal{C})$.

Bước quy nạp: Với giả thiết quy nạp là mỗi lớp $\mathcal{S}_{n-1}(\mathcal{C}')$ có không quá một phần tử $x \in X_{(n-1)}$ mà $v(x) \in \mathcal{S}_{n-1}(\mathcal{C}')$; chúng ta sẽ chứng minh rằng mỗi lớp $\mathcal{S}_n(\mathcal{C})$ cũng có không quá một phần tử $x \in X_{(n)}$ mà $v(x) \in \mathcal{S}_n(\mathcal{C})$ qua hai trường hợp sau:

a) Nếu tồn tại $x \in X_n$ và $v(x) \in \mathcal{S}_n(\mathcal{C})$, tức là $\mathcal{S}_n(x) = \mathcal{S}_n(\mathcal{C})$ thì theo Mệnh đề 4.3, x là phần tử duy nhất trong $X_{(n)}$ có $v(x) \in \mathcal{S}_n(\mathcal{C})$.

b) Nếu tồn tại $y \in X_{(n-1)}$ và $v(y) \in \mathcal{C}$ tức là $\mathcal{S}_n(y) = \mathcal{S}_n(\mathcal{C})$ thì theo Mệnh đề 4.3(b) với mọi $x \in X_n$, $v(x) \notin \mathcal{S}_n(y)$. Bây giờ xét những phần tử còn lại thuộc $X_{(n-1)}$. Vì các lớp tương đương của quan hệ $\mathcal{S}_n$ là phân hoạch trên một lớp của $\mathcal{S}_{n-1}$ nên ta luôn có $\mathcal{S}_n(\mathcal{C}) = \mathcal{S}_n(y) \subseteq \mathcal{S}_{n-1}(y) = \mathcal{S}_{n-1}(\mathcal{C}')$. Theo giả thiết quy nạp, đối với lớp $\mathcal{S}_{n-1}(\mathcal{C}')$ chỉ có duy nhất phần tử $y \in X_{(n-1)}$ sao cho $v(y) \in \mathcal{S}_{n-1}(\mathcal{C}')$. Vì vậy $\mathcal{S}_n(\mathcal{C})$ cũng chỉ có duy nhất phần tử $y \in X_{(n-1)}$ sao cho $v(y) \in \mathcal{S}_n(\mathcal{C})$. Hệ quả đã được chứng minh. ■

Với kí hiệu $X_{(k)} = X_1 \cup ... \cup X_k$, đặc biệt khi $k = k^*$ ta có $X_{(k)}$ chính là $LDom(A)$. Từ Hệ quả 4.2 ta suy ra một tính chất quan trọng về quan hệ bằng nhau mức k trong trường hợp này.

Mệnh đề 4.4. *Giả sử $LDom(A)$ hữu hạn, k^* là độ dài tối đa của dữ liệu trong $LDom(A)$ và fm là một độ đo tính mờ trên X. Với k tùy ý, $1 \leq k \leq k^*$ và bất kỳ $x, y \in X_{(k)}$, nếu $x =_k y$ thì $x = y$. Do đó, các quan hệ đối sánh mờ $=_k, \neq_k, \leq_k, \geq_k, <_k, >_k$ hạn chế trên $X_{(k)}$ trùng với các quan hệ $=, \neq, \leq, \geq, <, >$ thông thường được định nghĩa trên $X_{(k)}$.*

Chứng minh. Xét quan hệ tương tự mức k là $\mathcal{S}_k$. Với $x, y \in X_{(k)}$, nếu $x =_k y$ thì tồn tại một khoảng $\mathcal{S}_k(\mathcal{C})$ sao cho $\mathcal{S}_k(x) \subseteq \mathcal{S}_k(\mathcal{C})$ và $\mathcal{S}_k(y) \subseteq \mathcal{S}_k(\mathcal{C})$. Ta suy ra $v(x) \in \mathcal{S}_k(\mathcal{C})$ và $v(y) \in \mathcal{S}_k(\mathcal{C})$ và vì vậy $x = y$ theo Hệ quả 4.2. Kết hợp với Định nghĩa 4.4, dễ dàng suy ra các khẳng định còn lại. ■

4.3. Chuyển các truy vấn mờ cơ sở về truy vấn rõ

Vì có những điểm tương đồng giữa quan hệ đối sánh mờ mức k và quan hệ đối sánh thông thường nên chúng ta sẽ đưa ra cách chuyển đổi các truy vấn mờ cơ sở về các truy vấn rõ. Mỗi $x \in \underline{\mathcal{D}}_A$, lớp tương đương $\mathcal{S}_k(x)$ của quan hệ $=_k$ chứa lân cận tối thiểu $O_{min,k}(x)$ của x và mỗi lớp $\mathcal{S}_k(x)$ như vậy là một khoảng duy nhất. Kí hiệu $lf(\mathcal{S}_k(x))$ và $rt(\mathcal{S}_k(x))$ cho các điểm đầu cuối của đoạn $\mathcal{S}_k(x)$. Ngoại trừ trường hợp $lf(\mathcal{S}_k(x)) = 0$, ta quy ước $lf(\mathcal{S}_k(x)) \notin \mathcal{S}_k(x)$.

Nếu chỉ có dữ liệu kiểu 1 và kiểu 2 thì việc chuyển đổi khá đơn giản. Để tiện cho việc trình bày, kí hiệu $v^*(t[A])$ được sử dụng theo nghĩa sau: Nếu $t[A] \in D_A$ thì $v^*(t[A]) = t[A]$; ngoài ra nếu $t[A] \in LDom(A)$ thì $v^*(t[A]) = v_A(t[A])$.

Mệnh đề 4.5. *Trên $\mathcal{D}_A = D_A \cup LDom(A)$, ta có:*

a) $t[A] =_k x$ *khi và chỉ khi* $v^*(t[A]) \in \mathcal{S}_k(x)$;

b) $t[A] \neq_k x$ *khi và chỉ khi* $v^*(t[A]) \notin \mathcal{S}_k(x)$;

c) $t[A] \leq_k x$ *khi và chỉ khi* $t[A] =_k x$ *hoặc* $v^*(t[A]) \leq lf(\mathcal{S}_k(x))$;

d) $t[A] <_k x$ *khi và chỉ khi* $v^*(t[A]) \neq 0$ *và* $v^*(t[A]) \leq lf(\mathcal{S}_k(x))$ *hoặc* $v^*(t[A]) = 0$ *và* $v^*(t[A]) < lf(\mathcal{S}_k(x))$;

e) $t[A] \geq_k x$ *khi và chỉ khi* $t[A] =_k x$ *hoặc* $v^*(t[A]) \geq rt(\mathcal{S}_k(x))$;

f) $t[A] >_k x$ *khi và chỉ khi* $v^*(t[A]) > rt(\mathcal{S}_k(x))$.

Các khẳng định trên có thể kiểm chứng dễ dàng nhờ các Định nghĩa 4.3, 4.4 và Mệnh đề 4.1(b).

Mệnh đề 4.6. *Trên miền trị mở rộng $\underline{\mathcal{D}}_A$, nếu trong $t[A]$ và x có ít nhất một giá trị thuộc từ kiểu 3 đến kiểu 7 thì ta có:*

a) $t[A] =_k x$ *khi và chỉ khi* $t[A], x$ *đều tương thích mức* k *và* $O_{min,k}(t[A]) \in \mathcal{S}_k(x)$ *hoặc một trong hai giá trị* $t[A]$ *và* x *không tương thích mức* k*, nhưng đồng nhất nhau về kí hiệu (tức là* $t[A] = x$*);*

b) $t[A] \neq_k x$ *khi và chỉ khi* $t[A], x$ *đều tương thích mức* k *và* $v^*(t[A]) \notin \mathcal{S}_k(x)$ *hoặc một trong hai giá trị* $t[A], x$ *không tương thích mức* k*, và chúng không đồng nhất nhau về kí hiệu (tức là* $t[A] \neq x$*);*

c) $t[A] \leq_k x$ *khi và chỉ khi* $t[A] =_k x$ *hoặc* $t[A], x$ *đều tương thích mức* k *và* $O_{min,k}(t[A]) \leq lf(\mathcal{S}_k(x))$;

d) $t[A] <_k x$ *khi và chỉ khi* $t[A], x$ *đều tương thích mức* k *và* $O_{min,k}(t[A]) \leq lf(\mathcal{S}_k(x))$ *nếu* $lf(\mathcal{S}_k(x)) \neq 0$*, hoặc* $O_{min,k}(t[A]) < lf(\mathcal{S}_k(x))$ *nếu* $lf(\mathcal{S}_k(x)) = 0$;

e) $t[A] \geq_k x$ *khi và chỉ khi* $t[A] =_k x$ *hoặc* $t[A], x$ *đều tương thích mức* k *và* $O_{min,k}(t[A]) \geq rt(\mathcal{S}_k(x))$;

g) $t[A] >_k x$ *khi và chỉ khi* $t[A], x$ *đều tương thích mức* k *và* $O_{min,k}(t[A]) > rt(\mathcal{S}_k(x))$.

Để minh họa cho cách chuyển đổi truy vấn mờ sang truy vấn theo kiểu thông thường, ta xét ví dụ sau.

Ví dụ 4.1. Xét lược đồ quan hệ $R = \textit{Bảng-Lương} = \{\textit{Tên, Tuổi, Chức-vụ, Lương, Phụ-cấp}\}$. Giả sử một thể hiện của R được cho trong Bảng 4.1 (không cần để ý các dòng đầu), trong đó Lương và Phụ-cấp tính theo đơn vị triệu đồng. Bây giờ, ta xét truy vấn sau: **"Tìm những nhân viên trẻ có lương và phụ cấp khá cao"**. Điều kiện trong truy vấn trên có thể biểu diễn bởi biểu thức $t[\textbf{\textit{Tuổi}}] = \textbf{\textit{Trẻ}}$ **and** $t[\textbf{Lương}] \geq \textbf{\textit{Khá-cao}}$ **and** $t[\textbf{\textit{Phụ-cấp}}] \geq \textbf{\textit{Khá-cao}}$. Chúng ta sẽ tìm các bộ thỏa điều kiện khi chọn các mức $k = 1$ và $k = 2$.

Bảng 4.1. Quan hệ $R = \textit{Bảng-Lương}$

$k=1$	$\mathcal{S}_{1,Tuổi}(trẻ)$ =(21.528,31.23]		$\mathcal{S}_{1,Lương}(khá\text{-}cao)$ =(4.125,6.6]	$\mathcal{S}_{1,Phụ\text{-}cấp}(khá\text{-}cao)$ =(2.75,4.4]
$k=2$	$\mathcal{S}_{2,Tuổi}(trẻ)$ =(25.2324,27.91476]		$\mathcal{S}_{1,Lương}(khá\text{-}cao)$ =(5.75625,6.375]	$\mathcal{S}_{1,Phụ\text{-}cấp}(khá\text{-}cao)$ =(3.8375,4.25]
Tên	**Tuổi**	**Chức-vụ**	**Lương**	**Phụ-cấp**
Vĩnh	49	Giám đốc	*Rất-cao*(7.0950)	4.6
Linh	45	Trưởng phòng	6.2	4.2
Hùng**	27	Phó phòng	5.8	3.8
Thanh	35	Chuyên viên	*Khá-cao* (6.09375)	4.0
Tuyết*	31	Chuyên viên	4.7	*Cao* (3.65)
Việt	29	Chuyên viên	4.3	2.5
Tài	24	Chuyên viên	3.5	2.1

Giả sử $D_{Tuổi} = [18, 60]$, $D_{Lương} = [0, 7.5]$, $D_{Phụ\text{-}cấp} = [0, 5]$ và các tham số mờ gồm có: $fm(già) = 0.58$, $fm(trẻ) = 0.42$, $\mu(gần) = 0.27$, $\mu(ít) = 0.25$, $\mu(khá) = 0.28$ và $\mu(rất) = 0.20$. Vì vậy $\alpha = 0.52, \beta = 0.48$. Đối với các thuộc tính *Lương* và *Phụ-cấp* các tham số như nhau: $K = fm(c^-) = 0.40$, $\mu(ít) = 0.25$, $\mu(gần) = 0.30$, $\mu(khá) = 0.25$, $\mu(rất) = 0.20$, suy ra $\alpha = 0.55, \beta = 0.45$. Tính các giá trị có mặt trong R và phân hoạch mức 1, mức 2 cho các giá trị trong truy vấn:

(1) Thuộc tính *Tuổi*: theo Ví dụ 3.1, ta có $v_{Tuổi,r}(trẻ) = 26.4672$, $v_{Tuổi,r}(khá\text{-}trẻ) = 23.89882$

+ $\mathcal{S}_{1,Tuổi,r}(trẻ) = \mathfrak{I}_{2,Tuổi,r}(khá\text{-}trẻ) \cup \mathfrak{I}_{2,Tuổi,r}(gần\text{-}trẻ) = (26.4672 - 0.28 \times 0.42 \times 42, 26.4672 + 0.27 \times 0.42 \times 42] = (21.528, 31.23]$

+ $\mathcal{S}_{2,Tuổi,r}(trẻ) = \mathfrak{I}_{3,Tuổi,r}(ít\text{-}khá\text{-}trẻ) \cup \mathfrak{I}_{3,Tuổi,r}(rất\text{-}gần\text{-}trẻ) = (26.4672 - 0.25 \times 0.28 \times 0.42 \times 42, 26.4672 + 0.20 \times 0.27 \times 0.42 \times 42] = (25.2324, 27.91476]$.

(2) Thuộc tính *Lương*:

+ $v_{Lương,r}(cao) = [K + \alpha fm(cao)] \times 7.5 = [0.40 + 0.55 \times 0.60] \times 7.5 = 5.475$

+ $v_{Lương,r}(khá\text{-}cao) = v_{Lương,r}(cao) + \alpha \times \mu(khá) \times fm(cao) \times 7.5 = 5.475 + 0.55 \times 0.25 \times 0.60 \times 7.5 = 6.09375$

+ $v_{Lương,r}(rất\text{-}cao) = 7.5 - [(1-\alpha) \times \mu(rất) \times fm(cao)] \times 7.5 = 7.5 - 0.45 \times 0.20 \times 0.60 \times 7.5 = 7.095$

+ $\mathcal{S}_{1,Lương,r}(khá\text{-}cao) = \mathfrak{I}_{2,Lương,r}(gần\text{-}cao) \cup \mathfrak{I}_{2,Lương,r}(khá\text{-}cao)$
$= (5.475 - 0.30 \times 0.60 \times 7.5, 5.475 + 0.25 \times 0.60 \times 7.5] = (4.125, 6.6]$

+ $\mathcal{S}_{2,Lương,r}(khá\text{-}cao) = \mathfrak{I}_{3,Lương,r}(gần\text{-}khá\text{-}cao) \cup \mathfrak{I}_{3,Lương,r}(khá\text{-}khá\text{-}cao)$
$= (6.09375 - 0.30 \times 0.25 \times 0.60 \times 7.5, 6.09375 + 0.25 \times n0.25 \times 0.60 \times 7.5] = (5.75625, 6.375]$

(3) Thuộc tính *Phụ-cấp*:

$+\ v_{Phụ\text{-}cấp,r}(cao) = [K + \alpha \times fm(cao)] \times 5 = [0.40 + 0.55 \times 0.60] \times 5 = 3.65$
$+\ v_{Phụ\text{-}cấp,r}(khá\text{-}cao) = v_{Phụ\text{-}cấp,r}(cao) + \alpha \times \mu(khá) \times fm(cao) \times 5 = 3.65 + 0.55 \times 0.25 \times 0.60 \times 5 = 4.0625$
$+\ \mathcal{S}_{1,Phụ\text{-}cấp,r}(khá\text{-}cao) = \mathfrak{I}_{2,Phụ\text{-}cấp,r}(gần\text{-}cao) \cup \mathfrak{I}_{2,Phụ\text{-}cấp,r}(khá\text{-}cao)$
$= (3.65 - 0.30 \times 0.60 \times 5, 3.65 + 0.25 \times 0.60 \times 5] = (2.75, 4.4]$
$+\ \mathcal{S}_{2Phụ\text{-}cấp,r}(khá\text{-}cao) = \mathfrak{I}_{3,Phụ\text{-}cấp,r}(gần\text{-}khá\text{-}cao) \cup \mathfrak{I}_{3,Phụ\text{-}cấp,r}(khá\text{-}khá\text{-}cao)$
$= (4.0625 - 0.30 \times 0.25 \times 0.60 \times 5, 4.0625 + 0.25 \times 0.25 \times 0.60 \times 5] = (3.8375, 4.25]$

Như vậy, khi $k = 1$ sử dụng các quan hệ đối sánh đã nêu, so sánh với điều kiện trong truy vấn có hai bộ thỏa mãn ứng với Hùng** và Tuyết*. Khi $k = 2$, chỉ có duy nhất một bộ thỏa mãn là Hùng**.

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một mô hình mới cho cơ sở dữ liệu ngôn ngữ với thông tin mờ, không chính xác và có nhiều kiểu dữ liệu khác nhau dựa trên cấu trúc định lượng của ĐSGT. Quan điểm chính là biểu diễn dữ liệu bằng tập các khoảng trên không gian tham chiếu của miền trị các thuộc tính tương tự như cách biểu diễn qua lát cắt α của tập mờ. Tuy nhiên, nhờ cách biểu diễn các giá trị ngôn ngữ theo cú pháp và cấu trúc thứ tự trong ĐSGT nên việc xây dựng các quan hệ đối sánh được thực hiện dễ dàng hơn. Đồng thời phương pháp cũng có một số ưu điểm nhất định: biểu diễn các kiểu dữ liệu khác nhau một cách thống nhất; quan hệ tương tự được định nghĩa một cách mềm dẻo với nhiều mức khác nhau; các tham số được sử dụng trong phương pháp là ít và dễ xác định; quan hệ bằng nhau mức k giống như quan hệ bằng nhau thông thường nên có thể thao tác trên các truy vấn mờ như là truy vấn rõ. Phụ thuộc hàm, phụ thuộc dữ liệu trên mô hình này sẽ được tiếp tục nghiên cứu trong những công trình tiếp theo.

TÀI LIỆU THAM KHẢO

[1] S. Al-Hamouz and R. Biswas, Fuzzy functional dependencies in relational databases, *Intern. J. of Computational Cognition* (http://www.ijcc.us) Vol. **4** (1) (2006) 39–43.

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

[3] Billy P. Buckles, Extending the fuzzy database with fuzzy numbers, *Information Sciences* **34** (1984) 145–155.

[4] B.P. Buckles, F.E. Petry, A fuzzy representation of data for relational databases, *Fuzzy Sets and Systems* **7** (3) (1982) 213–226.

[5] B.P. Buckles, F.E. Petry, Fuzzy databases and their applications, In M. Gupta - E. Sanchez (Eds.), *Fuzzy Information and Decision Processes* (Vol. 2) North-Holland, Amsterdam, 1982 (361–371).

[6] B.P. Buckles, F.E. Petry, Uncertainty models in information and database systems, *Information Sciences* **11** (1985) 77–87.

[7] S.K. De, R. Biswas, A.R. Roy, On extended fuzzy relational database model with proximity relations, *Fuzzy Sets and Systems* **117** (2001) 195–201.

[8] Elke A. Rundensteiner, Lois W. Hawkes, and Wyllis Bandler, On nearness measures in fuzzy relational data models, *Intern. J. of Approx. Reasoning* **3** (1989) 267–298.

[9] Torsten Polle, Torsten Ripke, Klaus-Dieter Schewe (Eds.), *Fundamentals of Information Systems*, Kluwer Academic Publisher, 1999.

[10] 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).

[11] N.C. Ho, A Topological Completion of Refined Hedge Algebras and a Model of Fuzziness of Linguistic terms, *Fuzzy Sets and Systems* 158(4) (2007) 436–451.

[12] 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(4) (2007) 452–471.

[13] N.C. Ho, W. Wechler, Hedge algebras: An algebraic approach to structures of sets of linguistic domains of linguistic truth variable, *Fuzzy Sets and Systems* **35** (3) (1990) 281–293.

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

[15] Nguyen Cat Ho, Vu Nhu Lan, Le Xuan Viet, Optimal hedge-algebras-based controller: Design and Application, *Fuzzy Sets and Systems* 159 (2008) 968– 989.

[16] José Galindo, Angélica Urrutia, Mario Piattini, *Fuzzy Database: Modeling, Design and Implementation*, Idea Group Publishing, 2006 (Hershey - London -Melbourne - Singapore).

[17] H. Prade, C. Testemale, Generalizing database relational algebra for the treatment of incomplete/uncertain information and vague queries, *Inform. Sciences* **34** (1984) 115–143.

[18] S. Shenoi, A. Melton, Proximity relations in the fuzzy relational database model, *Fuzzy Sets and Systems* **5** (1) (1981) 31–46.

[19] S. Shenoi, A. Melton, An equivalence classes model of fuzzy relational databases, *Fuzzy Sets and Systems* **38** (1990) 153–170.

[20] M. Umano, Freedom-O: A fuzzy database system, *Fuzzy Information and Decision Processes*, M. Gupta, E. Sanchez (Eds.), North-Holland, Amsterdam, 1982 (339–347).

[21] M. Umano, Retrieval from fuzzy database by fuzzy relational algebra, *Fuzzy Information, Knowledge Representation And Decision Analysis*, M. Gupta, E. Sanchez (Eds.) New York: Pergamon Press, 1983 (1–6).

[22] N. Waraporn, and K. Porkaew, Null semantics for subqueries and atomic predicates, *IAENG International Journal of Computer Science* **35** (3) (2008) IJCS-35-3-08 (Advance online publication, 21 August 2008).

[23] Wei Yi Liu, Fuzzy data dependencies and implication of fuzzy data dependencies, *Fuzzy Sets and Systems* **92** (1997) 341–348.

[24] L. A. Zadeh, From Computing With Numbers To Computing With Words - From Manipulation Of Measurements To Manipulation Of Perceptions, *Int. J. Appl. Math. Comput. Sci.* **12** (3) (2002) 307–324.

Nhận bài ngày 21 - 12 - 2009