Tạp chí Tin học và Điều khiển học, T.24, S.4 (2008), 281–294

MỘT PHƯƠNG PHÁP XỬ LÝ TRUY VẤN TRONG CSDL MỜ TIẾP CẬN NGỮ NGHĨA LÂN CẬN CỦA ĐẠI SỐ GIA TỬ

NGUYỄN CÁT HỒ[1], NGUYỄN CÔNG HÀO[2]

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

Abstract. In this paper, we introduce a fuzzy database query language to manipulate of data in fuzzy database with data semantics on hedge algebras [4]. Data manipulation language is put forward for accord with a new model. Finally, we add quantifiers in natural languages in query for searching data which are rich and soft.

Tóm tắt. Trong bài báo này, chúng tôi giới thiệu ngôn ngữ truy vấn dùng để thao tác dữ liệu trên mô hình cơ sở dữ liệu mờ theo cách tiếp cận đại số gia tử [4]. Các ngôn ngữ thao tác dữ liệu được đề xuất phù hợp với mô hình mới. Cuối cùng, chúng tôi đưa thêm lượng từ ngôn ngữ vào câu truy vấn để việc thao tác dữ liệu được đa dạng và mềm dẻo.

1. MỞ ĐẦU

Ngôn ngữ truy vấn trong cơ sở dữ liệu (CSDL) mờ đã có nhiều tác giả quan tâm nghiên cứu và đã thu được một số kết quả trong thời gian qua [7 – 14]. Các cách xử lý truy vấn chủ yếu dựa trên cơ sở của cách tiếp cận lý thuyết tập mờ [9 – 13], như quan hệ tương tự [8]. Hầu hết, các tác giả đều xây dựng ngôn ngữ truy vấn với mong muốn thao tác mềm dẻo, "chính xác" với dữ liệu mờ bằng cách tập trung xây dựng các hàm thuộc, từ đó tùy theo ngữ nghĩa của dữ liệu để chọn các ngưỡng phù hợp khi thao tác dữ liệu. Tuy nhiên, mỗi ngôn ngữ truy vấn chỉ phù hợp với một mô hình CSDL mờ cụ thể mà không có ngôn ngữ nào tổng quát.

Dựa trên cách tiếp cận đại số gia tử (ĐSGT) một mô hình mới của CSDL mờ đã được xây dựng, 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 mô hình mới cho phép tìm kiếm, đánh giá ngữ nghĩa của thông tin không chắc chắn cũng như dữ liệu kinh điển một cách thống nhất trên cơ sở bảo đảm tính thuần nhất của kiểu dữ liệu trong xử lý ngữ nghĩa 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 được xem như là biến ngôn ngữ.

Với những ưu điểm của mô hình được đề xuất trong [4], trong bài báo này chúng tôi

nghiên cứu một phương pháp mới xử lý truy vấn mờ và xem xét việc đưa các lượng từ vào trong câu truy vấn để tìm kiếm dữ liệu.

Trong phần 2 chúng tôi trình bày một số khái niệm cơ bản liên quan đến ĐSGT và việc xây dựng mô hình CSDL mờ làm cơ sở cho các mục tiếp theo. Phần 3 trình bày các thuật toán xử lý trong truy vấn và các phương pháp đánh giá lượng từ trong câu truy vấn.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN

Để dễ theo dõi phương pháp xử lý ngữ nghĩa ngôn ngữ theo cách tiếp cận ĐSGT, chúng 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 cũng như độ tương tự mức k.

Cho một ĐSGT tuyến tính đầy đủ $\mathcal{AX} = (\boldsymbol{X}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \leq)$, trong đó $Dom(\mathcal{X}) = \mathcal{X}$ là miền các giá trị ngôn ngữ của thuộc tính ngôn ngữ $\mathcal{X}$ được sinh tự do từ tập các phần thử sinh $\boldsymbol{G} = \{\mathbf{1}, c^+, \boldsymbol{W}, c^-, \mathbf{0}\}$ bằng việc tác động tự do các phép toán một ngôi 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 $\boldsymbol{\Phi} x = infimum\boldsymbol{H}(x)$, trong đó $\boldsymbol{H}(x)$ là tập các phần từ sinh ra từ x, còn quan hệ $\leq$ 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 Thunhap là "Tổng thu nhập của công nhân trong một tháng", thì $Dom(Thunhap) = \{$high, low, very high, more high, possibly high, very low, possibly low, less low,...$\}$, $\boldsymbol{G} = \{\mathbf{1}, high, \boldsymbol{W}, low, \mathbf{0}\}$, $\boldsymbol{H} = \{very, more, possibly, less\}$ và $\leq$ một quan hệ thứ tự cảm sinh từ ngữ nghĩa của các từ trong $Dom(Thunhap)$, chẳng hạn ta có $veryhigh > high, morehigh > high, possiblyhigh < high, lesshigh < high, ...$

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

Ký hiệu $fm : \boldsymbol{X} \to [0, 1]$ là độ đo tính mờ của ĐSGT $\mathcal{AX}$. Khi đó,

Định nghĩa 2.1. Với mỗi $x \in \boldsymbol{X}$, độ dài của x được ký hiệu $|x|$ và xác định như sau:

(1) Nếu $x = c^+$ hoặc $x = c^-$ thì $|x| = 1$.

(2) Nếu $x = hx$ thì $|x| = 1 + |x|$, với mọi $h \in \boldsymbol{H}$.

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

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

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

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

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

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

2.1. Khoảng mờ của khái niệm mờ

Giả sử thuộc tính $\mathcal{X}$ 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 $[0, 1]$. Khi đó, tính chất (2) trong Mệnh đề 2.1 cho phép ta xây dựng hai khoảng mờ của hai khái niệm nguyên

thủy c^- và c^+, ký hiệu là $I(c^-)$ và $I(c^+)$ với độ dài tương ứng là $fm(c^-)$ và $fm(c^+)$ sao cho chúng tạo thành một phân hoạch của miền tham chiếu $[0,1]$ và $I(c^-)$ và $I(c^+)$ là đồng biến với c^- và c^+, tức là $c^- = c^+$ kéo theo $I(c^-) = I(c^+)$.

Một cách quy nạp, giả sử rằng với $\forall x \in \boldsymbol{X}_{k-1} = \{x \in \boldsymbol{X} : x$ có độ dài $|x| = k-1\}$, ta đã xây dựng được hệ các khoảng mờ $\{I(x) : x \in \boldsymbol{X}_{k-1}$ và $|I(x)| = fm(x)\}$ sao chúng là đồng biến và tạo thành một phân hoạch của đoạn $[0,1]$. Khi đó, trên mỗi khoảng mờ $I(x)$, độ dài $fm(x)$, của $x \in \boldsymbol{X}_{k-1}$, nhờ tính chất (4) trong Mệnh đề 2.1, ta có thể xây dựng được họ $\{I(h_i x) : q \le i \le p,\ i \ne 0,\ |I(h_i x)| = fm(h_i x)\}$ sao cho chúng là một phân hoạch của khoảng mờ $I(x)$. Có thể thấy họ $\{I(h_i x) : q \le i \le p,\ i \ne 0,\ |I(h_i x)| = fm(h_i x)$ và $x \in \boldsymbol{X}_{k-1}\} = \{I(y) : y \in \boldsymbol{X}_k$ và $|I(y)| = fm(y)\}$ là một phân hoạch của $[0,1]$. Các khoảng này gọi là các khoảng mờ mức k.

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

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

(2) $Sgn(h'hx) = 0$ nếu $hhx = hx$, còn ngược lại ta có

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

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

Định nghĩa 2.3. Giả sử $\mathcal{AX} = (\boldsymbol{X}, \boldsymbol{G}, \boldsymbol{H}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \le)$ 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) $\nu(W) = \kappa = fm(c^-),\ \nu(c^-) = \kappa - \alpha fm(c^-) = \beta fm(c^-),\ \nu(c^+) = \kappa + \alpha fm(c^+)$;

(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 $j, -q \le j \le p$ và $j \ne 0$;

(3) $\nu(\Phi c^-) = 0, \nu(\Sigma c^-) = \kappa = \nu(\Phi c^+), \nu(\Sigma c^+) = 1$, và với mọi $j, -q \le j \le p$ và $j \ne 0$, chúng ta có $\nu(\Phi h_j x) = \nu(x) + Sgn(h_j x)\{\sum_{i=sign(j)}^{j-1} \mu(h_i) fm(x)\}$ và $\nu(\Sigma h_j x) = \nu(x) +$

$Sgn(h_j x)\{\sum_{i=sign(j)}^{j} \mu(h_i) fm(x)\}$.

Nhận xét: Khoảng mờ $I(x)$ được xác định $I(x) = (\nu_A(\Phi x), \nu_A(\Sigma x)]$.

2.2. Độ tương tự mức k

Chúng 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 của lân cận mức k. Vì vậy ta định nghĩa độ tương tự mức k như sau: Chúng ta luôn luôn 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 từ độ dài k. Dựa trên các khoảng mờ mức k và các khoảng mờ mức k+1 chúng ta mô tả không hình thức việc xây dựng một phân hoạch của miền $[0, 1]$ như sau.

Với $k = 1$, các khoảng mờ mức 1 gồm $I(c^-)$ và $I(c^+)$. Các khoảng mờ mức 2 trên khoảng $I(c^-)$ là $I(h_pc^-) = I(h_{p-1}c^-) = ... = I(h_2c^-) = I(h_1c^-) = \nu_A(c^-) = I(h_{-1}c^-) = I(h_{-2}c^-) = ... = I(h_{-q+1}c^-) = I(h_{-q}c^-)$. Khi đó, ta xây dựng phân hoạch về độ tương tự mức 1 gồm các lớp tương đương sau: $S(\mathbf{0}) = I(h_pc^-); S(c^-) = I(c^-)\backslash[I(h_{-q}c^-) \cup I(h_pc^-)]; S(\boldsymbol{W}) = I(h_{-q}c^-) \cup I(h_{-q}c^+)$; tương tự ta có $S(c^+) = I(c^+)\backslash[I(h_{-q}c^+) \cup I(h_pc^+)]$ và $S(\mathbf{1}) = I(h_pc^+)$.

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, chẳng hạn, $I(h_ic^+) = (\nu_A(\Phi h_ic^+), \nu_A(\Sigma h_ic^+)]$ với hai khoảng mờ kề là $I(h_{i-1}c^+)$ và $I(h_{i+1}c^+)$ chúng ta sẽ có các lớp tương đương dạng sau: $S(h_ic^+) = I(h_ic^+)\backslash[I(h_ph_ic^+) \cup I(h_{-q}h_ic^+)], S(\Phi h_ic^+) = I(h_{-q}hi - 1c^+) \cup I(h_{-q}h_ic^+)$ và $S(\Phi h_ic^+) = I(h_ph_ic^+) \cup I(h_ph_ic^+)$, với i sao cho $-q \leq i \leq p$ và $i \neq 0$.

Bằng cách tương tự như vậy ta có thể xây dựng các phân hoạch các lớp tương tự mức k bất kỳ. Tuy nhiên trong thực tế ứng dụng chúng ta có thể giới hạn các gia tử tác động liên tiếp lên phần tử nguyên thủy c^- và c^+ là một số nguyên p nào đó. Các giá trị kinh điển và các giá trị mờ gọi là có độ tương tự mức k nếu các giá trị đại diện của chúng 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 mờ u chỉ và chỉ thuộc về một lớp tương tự, chẳng hạn đó là $S(x_i)$ và nó gọi là lân cận mức k của u và ký hiệu là $\Omega_k(u)$.

2.3. Cơ sở dữ liệu mờ

Xét một lược đồ CSDL $F_{DB} = \{U, R_1, R_2, ..., R_m; const\}$, 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 và được gọi là thuộc tính mờ. Tuy nhiên, trong bài báo này, chúng tôi chỉ xét những thuộc tính mờ mà miền trị của nó tồn tại một thứ tự tuyến tính. Những thuộc tính còn lại được gọi là thuộc tính kinh điển. Thuộc tính kinh điển A được gắn với một miền giá trị kinh điển, ký hiệu là D_A. Thuộc tính mờ 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. Một CSDL như vậy được gọi là CSDL mờ theo cách tiếp cận đại số gia tử.

2.4. Các quan hệ đối sánh trên miền trị thuộc tính

Định nghĩa 2.4. 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 là chúng bằng nhau mức k, nếu một trong các điều kiện sau xảy ra:

(1) Nếu $t[A_i], s[A_i] \in D_{Ai}$ thì $t[A_i] = s[A_i]$ hoặc là

(2) 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 $s[A_i] \in \Omega_k(t[A_i])$ hoặc là

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

Nếu điều kiện $t[A_i] =_{k(A_i)} s[A_i]$ không xảy ra ta có biểu thức $t[A_i] \neq_{k(A_i)} s[A_i]$.

Mệnh đề 2.2. *Quan hệ* $=_{k(A_i)}$ *là quan hệ tương đương trên* $[0, 1]$.

Dựa trên quan hệ tương đương $=_{k(A_i)}$, 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 $t[A_i] \in D_{Ai}$. 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, $\Omega_k(x)$ và $\Omega_k(y)$, ta sẽ viết $\Omega_k(x) < \Omega_k(y)$ khi $u < v$, với mọi $u \in \Omega_k(x)$ và mọi $v \in \Omega_k(y)$.

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

(1) Ta viết $t[A_i] \leq_k s[A_i]$, nếu $t[A_i] =_k s[A_i]$ hoặc là $\Omega_k(t[A_i]) < \Omega_k(s[A_i])$.

(2) Ta viết $t[A_i] <_k s[A_i]$, nếu $\Omega_k(t[A_i]) < \Omega_k(s[A_i])$.

(3) Ta viết $t[A_i] >_k s[A_i]$, nếu $\Omega_k(t[A_i]) > \Omega_k(s[A_i])$.

3. TRUY VẤN TRONG CƠ SỞ DỮ LIỆU MỜ

Dựa trên mô hình CSDL mờ trong [4], chúng tôi sẽ thiết kế một ngôn ngữ dùng để truy vấn dữ liệu trên mô hình đó. Trong phần này, chúng tôi sẽ đề xuất một phương pháp để xử lý câu truy vấn trong CSDL mờ mà chúng tôi gọi là SQL mờ và việc đưa các lượng từ tuyệt đối, lượng từ tỉ lệ vào trong câu truy vấn cũng được xem xét cụ thể.

3.1. Câu truy vấn SQL mờ

Các ngôn ngữ truy vấn trên mô hình CSDL quan hệ được nhiều tác giả quan tâm nghiên cứu và mở rộng trên mô hình CSDL mờ như: đại số quan hệ mờ, truy vấn SQL mờ. Vì thế, tương tự như trong CSDL quan hệ, cấu trúc của câu truy vấn SQL mờ được xem xét như sau: *select* ⟨ các thuộc tính ⟩ *from* ⟨các quan hệ⟩ where $\langle fc \rangle$, trong đó $\langle fc \rangle$ là điều kiện mờ hoặc liên kết các điều kiện mờ có sử dụng các phép toán tuyển và hội.

Như vậy, vấn đề quan trọng trong câu truy vấn SQL mờ chính là xác định giá trị chân lý của các $\langle fc \rangle$ và liên kết các giá trị chân lý đó.

Không mất tính tổng quát, chúng ta giả sử trong trường hợp đơn điều kiện, điều kiện mờ $\langle fc \rangle$ có dạng $A_i =_{k(A_i)} fvalue$, với $fvalue$ là giá trị mờ và A_i là thuộc tính mờ, khi đó ta có thể xây dựng một thuật toán như sau.

Thuật toán 3.1. Xử lý truy vấn SQL mờ trong trường hợp đơn điều kiện.

Vào : Quan hệ r xác định trên tập vũ trụ các thuộc tính $U = \{A_1, A_2, ..., A_n\}$.

Câu truy vấn dạng *select ...from... r* where $A_i =_{k(A_i)} fvalue$.

Ra : Quan hệ r_{result} thỏa mãn với mọi $t \in r_{result}$ ta có $t[A_i] =_{k(A_i)} fvalue$.

Phương pháp

Khởi tạo các giá trị

(1) Cho $G_{Ai} = \{\mathbf{0}, c^-_{A_i}, \boldsymbol{W}, c^+_{A_i}, \mathbf{1}\}$, $H_{Ai} = H^+_{Ai} \cup H^-_{Ai}$. Trong đó $H^+_{Ai} = \{h_1, h_2\}, H^-_{Ai} = \{h_3, h_4\}$, với $h_1 < h_2$ và $h_3 > h_4$. Chọn độ đo tính mờ cho các phần tử sinh và gia tử.

(2) $D_{Ai} = [\min_{Ai}, \max_{Ai}], \min_{Ai}, \max_{Ai}$: giá trị nhỏ nhất và lớn nhất miền trị kinh điển A_i.

(3) $LD_{Ai} = H_{Ai}(c^+) \cup H_{Ai}(c^-)$.

(4) $r_{result} = \emptyset$

Phân hoạch D_{Ai} thành các khoảng tương tự mức k.

(5) $k = 1$

(6) While $k = p$ do // mức tương tự lớn nhất $k = p$

(7) Xây dựng các khoảng tương tự mức $k : S_{Ai}(x_1), S_{Ai}(x_2), ..., S_{Ai}(x_m)$

(8) $k = k + 1$

Xác định lân cận mức k của $fvalue$

(9) if $fvalue \in S(x_i)$ then $\Omega_k(fvalue) = S(x_i)$

Duyệt các bộ t trong r để tìm các bộ thỏa mãn điều kiện $t[A_i] = k(A_i)fvalue$

(10) for each $t \in r$ do

(11) if $t[A_i] \in LD_{Ai}$ then

(12) Xác định lân cận mức k của $t[A_i]$ là $\Omega_k(t[A_i]) = S(x_j)$

(13) if $\Omega_k(t[A_i]) = \Omega_k(fvalue)$ then $r_{result} = r_{result} \cup t$

(14) elseif

(15) if $t[A_i] \in \Omega_k(fvalue)$ then $r_{result} = r_{result} \cup t$

(16) Return r_{result}

Ví dụ 3.1. Cho lược đồ quan hệ $U = \{STT, TENNV, NGHENGHIEP, TUOI, LUONG\}$ và quan hệ $Nhanvien$ được xác định như Bảng 3.1.

Bảng 3.1. Quan hệ Nhanvien

STT	TENNV	NGHENGHIEP	TUOI	LUONG
1	An	Giáo viên	45	rất cao
2	Bình	Kỹ sư	33	1100
3	Hà	Bác sĩ	rất khả năng trẻ	500
4	Hương	Y sĩ	36	700
5	Nhân	Giáo viên	46	1500
6	Thuỷ	Kiến trúc sư	26	khả năng cao
7	Thành	Y tá	trẻ	750
8	Xuân	Thư ký	21	thấp
9	Yến	Kỹ thuật viên	ít già	1125

Giả sử từ quan hệ $Nhanvien$ chúng ta muốn thực hiện câu truy vấn SQL mờ: Cho biết những cán bộ có tuổi khả năng trẻ. Sử dụng Thuật toán 3.1 ta có:

Bước (1)-(4): Cho $G_{TUOI} = \{\mathbf{0},$ trẻ, $W,$ già, $\mathbf{1}\}$, $D_{TUOI} = [0, 100]$, $H_{TUOI} = H^+_{TUOI} \cup H^-_{TUOI}$. Trong đó $H^+_{TUOI} = \{hơn,\ rất\}$, với $hơn < rất$ và $H^-_{TUOI} = \{ít,\ khả\ năng\}$, với $ít > khả\ năng$. Chọn $fm(già) = fm(c^+) = 0,35$, $fm(trẻ) = fm(c^-) = 0,65$, $\mu(khả\ năng) = 0{,}25$, $\mu(\text{ít}) = 0,20$, $\mu(hơn) = 0,15$ và $\mu(rất) = 0,40$. $LD_{TUOI} = H_{TUOI}(trẻ) \cup H_{TUOI}(già)$, $r_{result} = \emptyset$.

Bước (5)-(8): Vì $|\ khả\ năng\ trẻ\ | = 2$ nên ta chỉ cần đi xây dựng các khoảng tương tự mức 2. Phân hoạch đoạn $[0, 100]$ thành các khoảng tương tự mức 2: $fm(rất\ rất\ trẻ) \times 100 = 0,40 \times 0,40 \times 0,65 \times 100 = 10,4$. Vậy $S(0) \times 100 = [0; 10,4]$.

$(fm(\textit{hơn rất trẻ}) + fm(\textit{khả năng rất trẻ})) \times 100 = (0,15 \times 0,40 \times 0,65 + 0,25 \times 0,40 \times 0,65)100 = 10,4$ và $S(\textit{rất trẻ}) \times 100 = (10,4; 20,8]$.

$(fm(\textit{ít rất trẻ}) + fm(\textit{rất hơn trẻ})) \times 100 = (0,20 \times 0,40 \times 0,65 + 0,40 \times 0,15 \times 0,65) \times 100 = 9,1$.

$(fm(\textit{hơn hơn trẻ}) + fm(\textit{khả năng hơn trẻ})) \times 100 = (0,15 \times 0,15 \times 0,65 + 0,25 \times 0,15 \times 0,65) \times 100 = 3,9$ và $S(\textit{hơn trẻ}) \times 100 = (29,9; 33,8]$.

$(fm(\textit{ít hơn trẻ}) + fm(\textit{rất khả năng trẻ})) \times 100 = (0,20 \times 0,15 \times 0,65 + 0,40 \times 0,25 \times 0,65) \times 100 = 8,45$.

$(fm(\textit{hơn khả năng trẻ}) + fm(\textit{khả năng khả năng trẻ})) \times 100 = (0,15 \times 0,25 \times 0,65 + 0,25 \times 0,25 \times 0,65) \times 100 = 6,5$ và $S(\textit{khả năng trẻ}) \times 100 = (42,25; 48,75]$.

$(fm(\textit{ít khá trẻ}) + fm(\textit{rất ít trẻ})) \times 100 = (0,20 \times 0,25 \times 0,65 + 0,40 \times 0,20 \times 0,65) \times 100 = 8,45$.

$(fm(\textit{hơn ít trẻ}) + fm(\textit{khả năng ít trẻ})) \times 100 = (0,15 \times 0,20 \times 0,65 + 0,25 \times 0,20 \times 0,65) \times 100 = 5,2$ và $S(\textit{ít trẻ}) \times 100 = (57,2; 62,4]$. Tương tự, chúng ta tính được $S(\boldsymbol{W}), S(\textit{ít già}), S(\textit{khả năng già}), S(\textit{hơn già}), S(\textit{rất già}), S(1)$.

Như vậy, các khoảng tương tự mức 2 là: $S(\mathbf{0}), S(\textit{rất trẻ}), S(\textit{hơn trẻ}), S(\textit{khả năng trẻ}), S(\textit{ít trẻ}), S(\boldsymbol{W}), S(\textit{ít già}), S(\textit{khả năng già}), S(\textit{hơn già}), S(\textit{rất già}), S(\mathbf{1})$.

Bước (9): Xác định lân cận mức 2 của *khả năng trẻ*. Ta có *khả năng trẻ* $\in S(\textit{khả năng trẻ})$ nên lân cận mức 2 của *khả năng trẻ* là $\Omega_2(\textit{khả năng trẻ}) = S(\textit{khả năng trẻ}) = (42,25; 48,75]$.

Bước (10)-(15): Ta thấy trong quan hệ $Nhanvien, t_1[TUOI](\Omega_2(\textit{khả năng trẻ}), t_5[TUOI](\Omega_2(\textit{khả năng trẻ})$ và $\Omega_2(t_3[TUOI]) = \Omega_2(\textit{rất khả năng trẻ}) = \Omega_2(\textit{khả năng trẻ})$.

Bước (16): Vậy $r_{result} = \{t_1, t_3, t_5\}$.

Hay câu truy vấn SQL mờ $select * from Nhanvien\ where\ TUOI =_{2(TUOI)}$ *khả năng trẻ* cho kết quả sau

Bảng 3.2. Kết quả thực hiện truy vấn sử dụng Thuật toán 3.1

STT	TENNV	NGHENGHIEP	TUOI	LUONG
1	An	Giáo viên	45	rất cao
3	Hà	Bác sĩ	*rất khả năng trẻ*	500
5	Nhân	Giáo viên	46	1500

Thuật toán 3.2. Xử lý truy vấn SQL mờ trong trường hợp đa điều kiện

Vào: Quan hệ r xác định trên tập vũ trụ các thuộc tính $U = \{A_1, A_2, ..., A_n\}$.

Câu truy vấn dạng $select ... from r\ where A_i =_{k(Ai)} fvalue_i \Psi A_j =_{k(Aj)} fvalue_j$, trong đó Ψ là phép toán *and* hoặc *or*.

Ra : Quan hệ r_{result} thỏa mãn với mọi $t \in r_{result}$ ta có $t[A_i] =_{k(Ai)} fvalue_i \Psi t[A_j] =_{k(Aj)} fvalue_j$.

Phương pháp

Khởi tạo các giá trị

(1) Cho $G_{Ai} = \{\mathbf{0}, c^-_{Ai}, W, c^+Ai, \mathbf{1}\}, H_{Ai} = H^+_{Ai} \cup H^-_{Ai}$. Trong đó $H^+_{Ai} = \{h_1, h_2\}, H^-_{Ai} = \{h_3, h_4\}$, với $h_1 < h_2$ và $h_3 > h_4$. Chọn độ đo tính mờ cho các phần tử sinh và gia tử.

(2) Cho $G_{Aj} = \{\mathbf{0}, c^-Aj, \boldsymbol{W}, c^+_{Aj}, \mathbf{1}\}$, $H_{Aj} = H^+_{Aj} \cup H^-_{Aj}$. Trong đó $H^+_{Aj} = \{h_1, h_2\}$, $H^-_{Aj} = \{h_3, h_4\}$, với $h_1 < h_2$ và $h_3 > h_4$. Chọn độ đo tính mờ cho các phần tử sinh và gia tử.

(3) $D_{Ai} = [\min_{Ai}, \max_{Ai}], \min_{Ai}, \max_{Ai}$: giá trị nhỏ nhất và lớn nhất miền trị kinh điển A_i.
(4) $D_{Aj} = [\min_{Aj}, maxAj], \min_{Aj}, \max_{Aj}$: giá trị nhỏ nhất và lớn nhất miền trị kinh điển A_j.
(5) $LD_{Ai} = H_{Ai}(c^+_{Ai}) \cup H_{Ai}(c^-_{Ai})$.
(6) $LD_{Aj} = H_{Aj}(c^+_{Aj}) \cup H_{Aj}(c^-_{Aj})$.
(7) $r_{result} = \emptyset$
Phân hoạch D_{Ai} và D_{Aj} thành các khoảng tương tự mức k.
(8) $k = 1$
(9) While $k \leq p$ do // mức tương tự lớn nhất $k = p$
(10) Xây dựng các khoảng tương tự mức $k : S_{Ai}(x_1), S_{Ai}(x_2), ..., S_{Ai}(x_m)$
(11) Xây dựng các khoảng tương tự mức $k : S_{Ai}(y_1), S_{Ai}(y_2), ..., S_{Ai}(y_m)$
(12) $k = k + 1$
Xác định lân cận mức k của $fvalue_i$ và $fvalue_j$
(13) if $fvalue_i \in S_{Ai}(x_i)$ then $\Omega_k(fvalue_i) = S_{Ai}(x_i)$
(14) if $fvalue_j \in S_{Aj}(y_i)$ then $\Omega_k(fvalue_j) = S_{Aj}(y_i)$
Duyệt các bộ t trong r để tìm các bộ thỏa mãn điều kiện $t[A_i] =_{k(Ai)} fvalue_i \Psi t[Aj] =_{k(Aj)} fvalue_j$
(15) for each $t \in r$ do
(16) if $t[A_i] \in LD_{Ai}$ then
(17) Xác định lân cận mức k của $t[A_i]$ là $\Omega_k(t[A_i]) = S_{Ai}(x_i)$
(18) if $t[Aj] \in LD_{Aj}$ then
(19) Xác định lân cận mức k của $t[A_j]$ là $\Omega_k(t[A_j]) = S_{Aj}(y_i)$
Trường hợp Ψ là phép toán *and*
(20) if $\{\Omega_k(t[A_i]) = \Omega_k(fvalue_i)\}$ *and* $\{\Omega_k(t[A_j) = \Omega_k(fvalue_j\}$ then
$r_{result} = r_{result} \cup t$
(21) if $t[A_i] \in \Omega_k(fvalue_i)$ *and* $t[A_j] \in \Omega_k(fvalue_j)$ then $r_{result} = r_{result} \cup t$
Trường hợp Ψ là phép toán *or*
(22) if $\{\Omega_k(t[A_i]) = \Omega_k(fvalue_i)\}$ *or* $\{\Omega_k(t[A_j) = \Omega_k(fvalue_j\}$ then
$r_{result} = r_{result} \cup t$
(23) if $\{\Omega_k(t[A_i]) = \Omega_k(fvalue_i)\}$ *or* $\{\Omega_k(t[A_j) = \Omega_k(fvalue_j\}$ then
$r_{result} = r_{result} \cup t$
(24) Return r_{result}

Ví dụ 3.2. Sử dụng quan hệ *Nhanvien* trong Ví dụ 3.1. Hãy cho biết những cán bộ Tuổi khả năng trẻ và có Lương rất cao. Sử dụng Thuật toán 3.2 ta có:

Bước (1)-(7): Cho $G_{TUOI} = \{\mathbf{0}, trẻ, \mathbf{W}, già, \mathbf{1}\}, D_{TUOI} = [0; 100], H_{TUOI} = H^+_{TUOI} \cup H^-_{TUOI}$, trong đó $H^+_{TUOI} = \{hơn,\ rất\}$, với $hơn < rất$ và $H^-_{TUOI} = \{ít,\ khả\ năngt\}$, với $ít > khả\ năng$. Chọn $fm(già) = fm(c^+) = 0,35,\ fm(trẻ) = fm(c^-) = 0,65,\ \mu(khả\ năng) = 0,25,\ \mu(ít) = 0,20,\ \mu(hơn) = 0,15$ và $\mu(rất) = 0,40$. $LD_{TUOI} = H_{TUOI}(trẻ) \cup H_{TUOI}(già)$.

Cho $G_{LUONG} = \{\mathbf{0}, thấp, \mathbf{W}, cao, \mathbf{1}\}$ và $D_{LUONG} = [400; 1600],\ H_{LUONG} = H^+_{LUONG} \cup H^-_{LUONG}$, trong đó $H^+_{LUONG} = \{hơn,\ rất\}$, với $hơn < rất$ và $H^-_{LUONG} = \{ít,\ khả\ năngt\}$, với

ít > *khả năng*. Chọn $fm(cao) = fm(c^+) = 0,6,\ fm(thấp) = fm(c^-) = 0,4,\ \mu(khả\ năng) = 0,15, \mu(ít) = 0,25, \mu(hơn) = 0,25$ và $\mu(rất) = 0,35$.

$LD_{LUONG} = H_{LUONG}(cao) \cup H_{LUONG}(thấp),\ r_{result} = \emptyset$.

Bước (8)-(12): Vì $|khả\ năng\ trẻ| = 2$ và $|rất\ cao| = 2$ nên ta chỉ cần đi xây dựng các khoảng tương tự mức 2.

Đối với thuộc tính $TUOI$: Ta phân hoạch đoạn $[0; 100]$ thành các khoảng tương tự mức 2, theo Ví dụ 3.1 ta có: $S(\mathbf{0}), S(rất\ trẻ), S(hơn\ trẻ), S(khả\ năng\ trẻ), S(ít\ trẻ), S(\boldsymbol{W}), S(ít\ già), S(khả\ năng\ già), S(hơn\ già), S(rất\ già), S(\mathbf{1})$.

Đối với thuộc tính $LUONG$: Ta phân hoạch đoạn $[400; 1600]$ thành các khoảng tương tự mức 2 : $fm(rất\ rất\ cao) \times 1200 = 0,35 \times 0,35 \times 0,6 \times 1200 = 88,2$. Vậy $S(\mathbf{1}) \times 1200 = (1511,8; 1600]$.

$(fm(hơn\ rất\ cao) + fm(khả\ năng\ rất\ cao)) \times 1200 = (0,25 \times 0,35 \times 0,6 + 0,15 \times 0,35 \times 0,6) \times 1200 = 100,8$ và $S(rất\ cao) \times 1200 = (1411; 1511.8]$.

$(fm(ít\ rất\ cao) + fm(rất\ hơn\ cao)) \times 1200 = (0,25 \times 0,35 \times 0,6 + 0,35 \times 0,25x0,6) \times 1200 = 126$.

$(fm(khả\ năng\ hơn\ cao) + fm(hơn\ hơn\ cao)) \times 1200 = (0,15 \times 0,25 \times 0,6 + 0,25 \times 0,25 \times 0,6) \times 1200 = 72$ và $S(hơn\ cao) \times 1200 = (1213, 1285]$.

$(fm(ít\ hơn\ cao) + fm(rất\ khả\ năng\ cao)) \times 1200 = (0,25 \times 0,25 \times 0,6 + 0,35 \times 0,15 \times 0,6) \times 1200 = 82,8$.

$(fm(hơn\ khả\ năng\ cao) + fm(khả\ năng\ khả\ năng\ cao)) \times 1200 = (0,25 \times 0,15 \times 0,6 + 0,15 \times 0,15 \times 0,6) \times 1200 = 43,2$ và $S(khả\ năng\ cao) \times 1200 = (1087; 1130,2]$.

$(fm(ít\ khả\ năng\ cao) + fm(rất\ ít\ cao)) \times 1200 = (0,25 \times 0,15 \times 0,6 + 0,35 \times 0,25 \times 0,6) \times 1200 = 90$.

$(fm(hơn\ ít\ cao) + fm(khả\ năng\ ít\ cao)) \times 1200 = (0,25 \times 0,25 \times 0,6 + 0,15 \times 0,25 \times 0,6) \times 1200 = 72$ và $S(ít\ cao) \times 1200 = (925; 997]$. Tương tự, chúng ta tính được $S(\boldsymbol{W}), S(ít\ thấp)$,

$S(khả\ năng\ thấp), S(hơn\ thấp), S(rất\ thấp), S(\mathbf{0})$. Như vậy, các khoảng tương tự mức 2 là: $S(\mathbf{0}), S(rất\ thấp), S(hơn\ thấp), S(khả\ năng\ thấp), S(ít\ thấp), S(\boldsymbol{W}), S(ít\ cao), S(khả\ năng\ cao), S(hơn\ cao), S(rất\ cao), S(\mathbf{1})$.

Bước (13)-(14): Xác định lân cận mức 2 của khả năng trẻ và *rất cao*. Ta có lân cận mức 2 của khả năng trẻ là $\Omega_2(khả\ năng\ trẻ) = S(khả\ năng\ trẻ) = (42,25; 48,75]$ và lân cận mức 2 của *rất cao* là $\Omega_2(rất\ cao) = S(rất\ cao) = (1411; 1511,8]$.

Bước (15)-(19): Theo Ví dụ 3.1, điều kiện $TUOI =_{2(TUOI)} khả\ năng\ trẻ$ ta có $r_{result} = \{t_1, t_3, t_5\}$. Xét điều kiện $LUONG =_{2(LUONG)} rất\ cao$, trong quan hệ $Nhanvien$, ta có lân cận mức 2 của $t_1[LUONG] = \Omega_2(t_1[LUONG]) = \Omega_2(rất\ cao)$, $t_5[LUONG] \in \Omega_2(rất\ cao)$. Do đó ta có $r_{result} = \{t_1, t_5\}$.

Bước (20)-(23): Vì Ψ là phép toán and nên kết hợp điều kiện $TUOI =_{(TUOI)} khả\ năng\ trẻ$ và $LUONG =_{2(LUONG)} rất\ cao$ ta có $r_{result} = \{t_1, t_5\}$.

Bước (24): Kết quả $r_{result} = \{t_1, t_5\}$.

Vậy, câu truy vấn SQL mờ $select * from\ Nhanvien\ where\ TUOI =_{(TUOI)} khả\ năng\ trẻ$ và $LUONG =_{2(LUONG)} rất\ cao$ cho kết quả sau:

Bảng 3.3. Kết quả thực hiện truy vấn sử dụng Thuật toán 3.2

STT	TENNV	NGHENGHIEP	TUOI	LUONG
1	An	Giáo viên	45	*rất cao*
5	Nhân	Giáo viên	46	1500

Để tìm kiếm mềm dẽo hơn, đôi khi người ta đưa vào câu truy vấn các từ như *hầu hết, một vài...* hay còn gọi là lượng từ. Trong phần tiếp theo, chúng tôi sẽ nghiên cứu về các dạng câu truy vấn như vậy.

3.2. Lượng từ ngôn ngữ

Để khai thác dữ liệu trên mô hình CSDL mờ nhiều tác giả mở rộng những ngôn ngữ hỏi trên mô hình quan hệ như đại số quan hệ, ngôn ngữ SQL... Điểm mở rộng đó chính là sử dụng các điều kiện mờ, chẳng hạn như "*tìm những cán bộ trẻ có nhiều công trình khoa học công bố trên tạp chí quốc tế có uy tín*", "*cho biết các mặt hàng bán trong siêu thị thu được lợi nhuận khá lớn*"....Việc xử lý các câu hỏi dạng như vậy chúng ta chỉ cần tìm những bộ dữ liệu "*thỏa mãn*" những điều kiện mờ. Tuy nhiên, khi gặp những yêu cầu như "*cho biết ít nhất 5 cán bộ trẻ có nhiều công trình khoa học công bố trên tạp chí quốc tế có uy tín*", "*cho biết một vài mặt hàng bán trong siêu thị thu được lợi nhuận khá lớn*"... thì vấn đề xử lý câu hỏi là phức tạp, bởi vì, ngoài việc tìm những bộ dữ liệu "*thỏa mãn*" những điều kiện mờ còn phụ thuộc vào các lượng từ "*ít nhất 5*" và "*một vài*". Zadeh [9] chia lượng từ ngôn ngữ thành hai loại đó là: lượng từ tuyệt đối (absolute quantifier) và lượng từ tỉ lệ (proportional quantifier). Lượng từ tuyệt đối thường dùng trong các mệnh đề có số lượng xác định như "*ít nhất 5*", "*nhiều hơn 3*".... Lượng từ tỉ lệ thể hiện những số lượng phụ thuộc vào số lượng tập các đối tượng đang xử lý, chẳng hạn như "*hầu hết*", "*một vài*"...

Vì vậy, đưa lượng từ và sử lý lượng tù trong câu truy vấn SQL mờ là vấn đề cần quan tâm giải quyết.

3.3. Đưa lượng từ ngôn ngữ vào câu truy vấn SQL mờ

Để đáp ứng yêu cầu thao tác dữ liệu trong mô hình CSDL mờ, trong phần này, chúng tôi nghiên cứu, xây dựng phương pháp đánh giá lượng từ tương đối và lượng từ tuyệt đối và việc đưa lượng từ như vậy vào câu truy vấn SQL mờ như thế nào?

Do truy vấn sử dụng lượng từ có thể xem là một sự mở rộng của truy vấn mờ cho nên một câu truy vấn SQL mờ sử dụng lượng từ có thể tổng quát dạng: *select* ⟨ các thuộc tính ⟩ *from* ⟨ các quan hệ⟩ *where* $Q(fc_1, fc_2, ...fc_n)$, trong đó Q là lượng từ và $fc_1, fc_2, ...fc_n$ là các điều kiện mờ. Chẳng hạn như trong quan hệ *Nhanvien* ở Ví dụ 3.1, tìm "*ít nhất một nữa*" nhân viên *Tuổi trẻ* và có *Lương thấp.* Khi đó câu truy vấn có dạng: *select * from Nhanvien where "ít nhất một nữa"* ($TUOI =_{1(TUOI)}$ *trẻ* và $LUONG =_{1(LUONG)}$ *cao*).

Không mất tính tổng quát, chúng ta giả sử các điều kiện mờ fc_i trong câu truy vấn có dạng $A_i =_{k(Ai)} fvalue_i$, trong đó A_i là thuộc tính mờ và $fvalue_i$ là giá trị mờ, phép toán liên kết các điều kiện là phép *and* hoặc *or*. Do đó, điều kiện $Q(fc_1, fc_2, ...fc_n)$ có dạng: $Q(A_1 =_{k(A1)} fvalue_1 \Psi A_2 =_{k(A2)} fvalue_2 ...\Psi A_1 =_{k(An)} fvalue_n)$, trong đó Ψ là phép *and* hoặc *or*.

3.3.1. Xây dựng phương pháp đánh giá lượng từ trong truy vấn

Để xây dựng phương pháp đánh giá lượng từ trong truy vấn, trước hết chúng ta xác định giá trị chân lý của các điều kiện mờ $(fc_1, fc_2, ...fc_n)$ đối với quan hệ tham gia truy vấn. Có nghĩa là tìm những bộ dữ liệu t thuộc quan hệ tham gia truy vấn thỏa mãn điều kiện mờ $(fc_1, fc_2, ...fc_n)$, đây chính là các bước thực hiện truy vấn SQL mờ trong Mục 3.1. Tiếp theo, đánh giá lượng từ trong câu truy vấn dựa vào những bộ dữ liệu vừa tìm được so với số bộ dữ liệu của quan hệ ban đầu tham gia truy vấn.

Gọi $D_r = [0..\|r\|]$, với $\|r\|$ là số bộ dữ liệu trong quan hệ r. Chúng ta có thể chia lượng từ thành hai trường hợp:

a. Trường hợp $\boldsymbol{Q}$ *là lượng từ tuyệt đối*: Ký hiệu $\|Q\|$ là số lượng xác định của lượng từ Q.

Nếu Q đơn điệu tăng: Ta xây dựng một hàm $f_Q^A : D_r \to \{0, 1\}$ sao cho: $\forall x \in D_r, f_Q^A(x) = 1$ nếu $x \geq \|Q\|$ và $f_Q^A(x) = 0$ nếu ngược lại.

Nếu Q đơn điệu giảm: Ta xây dựng một hàm $f_Q^D : D_r \to \{0, 1\}$ sao cho: $\forall x \in D_r, f_Q^D(x) = 1$ nếu $x = \|Q\|$ và $f_Q^D(x) = 0$ nếu ngược lại.

b. Trường hợp $\boldsymbol{Q}$ *là lượng từ tỷ lệ*: Khi ta nói hầu hết các bộ dữ liệu t trong r thỏa mãn điều kiện $(fc_1, fc_2, ...fc_n)$, có nghĩa là tổng số bộ dữ liệu t phải xấp xỉ $\|r\|$. Hoặc trong trường hợp khác, chỉ một ít các bộ dữ liệu t trong r thỏa mãn điều kiện $(fc_1, fc_2, ...fc_n)$, có nghĩa là tổng số bộ dữ liệu t phải xấp xỉ $1/\|r\|$. Hay một giả thiết ta thường gặp đó là *khoảng một nửa* các bộ dữ liệu t trong r thỏa mãn điều kiện $(fc_1, fc_2, ...fc_n)$, khi đó chắc chắn rằng tổng số bộ dữ liệu t phải là xấp xỉ của $\|r\|/2$.

Điều này gợi ý cho chúng ta có thể đánh giá lượng từ tỉ lệ dựa trên sự phân hoạch của $[0..\|r\|]$. Theo Mục 2.1 để chuẩn hóa $[0..\|r\|]$, nhờ một phép biến đổi tuyến tính, ta giả thiết mọi miền $D_r = [0..\|r\|]$ như vậy đều là khoảng $[0, 1]$. Khi đó ta xây dựng hai khoảng mờ của hai khái niệm nguyên thủy nhỏ và lớn, ký hiệu là $I(nhỏ)$ và $I(lớn)$ với độ dài tương ứng là $fm(nhỏ)$ và $fm(lớn)$ sao cho chúng tạo thành một phân hoạch của miền tham chiếu $[0, 1]$. Tiếp đến, đi xây dựng các lớp tương đương $S(\mathbf{1}), S(lớn), S(\boldsymbol{W}), S(nhỏ), S(\mathbf{0})$ dựa vào độ đo tính mờ của các gia tử và các khái niệm nguyên thủy.

Do đó, nếu gọi $\|r_1\|, \|r_2\|$ tương ứng là tổng số bộ dữ liệu t trong r thỏa mãn điều kiện $(fc_1, fc_2, ...fc_n)$ với lượng từ hầu hết và một ít thì $\|r_1\| \in S(\mathbf{1}) \times \|r\|$ và $\|r_2\| \in S(\mathbf{0}) \times \|r\|$.

Như vậy, ta có thể khẳng định rằng tổng số bộ dữ liệu t trong r thỏa mãn điều kiện $(fc_1, fc_2, ...fc_n)$ áp dụng với lượng từ Q được ký hiệu $\|r_Q\|$, khi đó: $\|r_Q\| \in S(\mathbf{1}) \times \|r\|$ hoặc $\|r_Q\| \in S(lớn) \times \|r\|$, hoặc $\|r_Q\| \in S(\boldsymbol{W}) \times \|r\|$, hoặc $\|r_Q\| \in S(nhỏ) \times \|r\|$, hoặc $\|r_Q\| \in S(\mathbf{0}) \times \|r\|$ hay nói cách khác: $\|r_Q\|/\|r\|$ phải thuộc về một trong các khoảng: $S(\mathbf{1}), S(lớn), S(\boldsymbol{W}), S(nhỏ), S(\mathbf{0})$.

Sau khi đã xây dựng phương pháp đánh giá lượng từ trong truy vấn, Thuật toán 3.3 được đề xuất nhằm ứng dụng cho việc xử lý câu hỏi có chứa lượng từ. Vì số lượng các lượng từ không nhiều nên ở đây sẽ liệt kê lượng từ tương đối QABS và lượng từ tuyệt đối QPRO trong phần đầu vào của thuật toán.

3.3.2. Thuật toán xử lý lượng từ trong câu truy vấn SQL mờ

Thuật toán 3.3. Xử lý lượng từ trong truy vấn

Vào : Quan hệ r xác định trên tập vũ trụ các thuộc tính $U = \{A_1, A_2, ..., A_n\}$.

Câu truy vấn $select * from\ r\ where\ Q(A_1 =_{k(A1)}\ fvalue_1 \Psi A_2 =_{k(A2)}\ fvalue_2 ... \Psi A_n =_{k(An)}$

$fvalue_n$), trong đó Ψ là phép *and* hoặc *or*, $Q \in Q_{ABS} \cup Q_{PRO}$ với
$Q_{ABS} = \{$*một ít, khoảng một nữa, hầu hết, với mọi*$\}$, $Q_{PRO} = \{$*ít nhất m, nhiều nhất m*$\}$.
Ra : Một quan hệ $r_{resultQ}$ chứa những bộ dữ liệu t thỏa mãn điều kiện $Q(A_1 =_{k(A1)} fvalue_1 \Psi A_2 =_{k(A2)} fvalue_2 ... \Psi A_1 =_{k(An)} fvalue_n)$.
Phương pháp
(1) $r_{resultQ} = \emptyset$
(2) Sử dụng Thuật toán 3.1 trong trường hợp đơn điều kiện và Thuật toán 3.2 trong trường hợp đa điều kiện ta có kết quả là quan hệ r_{result}.
(3) if $Q \in Q_{ABS}$ then
(4) if $f_Q^A(|r_{result}|) = 1$ or $f_Q^D(|r_{result}|) = 1$ then $r_{resultQ} = r_{resul}$
(5) elseif
(6) if $Q \in Q_{PRO}$ then
(7) Xây dựng các khoảng: $S(\mathbf{1}), S(lớn), S(\boldsymbol{W}), S(nhỏ), S(\mathbf{0})$
(8) Case Q of
(9) "*Một ít*" : if $(|r_{result}|/|r|) \in S(\mathbf{0})$ then $r_{resultQ} = r_{resul}$
(10) "*Khoảng một nữa*" : if $(|r_{result}|/|r|) \in S(\boldsymbol{W})$ then $r_{resultQ} = r_{resul}$
(11) "*Hầu hết*" : if $(|r_{result}|/|r|) \in S(\mathbf{1})$ then $r_{resultQ} = r_{resul}$
(12) "*Với mọi*" : if $(|r_{result}|/|r|) = 1$ then $r_{resultQ} = r_{resul}$
(13) Return $r_{resultQ}$

Ví dụ 3.3. Sử dụng quan hệ Nhanvien trong Ví dụ 3.1

(i) Cho biết *ít nhất* 5 nhân viên có *Tuổi khả năng trẻ*, khi đó câu truy vấn SQL mờ có dạng: $select * from\ Nhanvien\ where\ ít\ nhất 5(TUOI =_{2(LUONG)} khả\ năng\ trẻ)$.

Bước (1): $r_{resultQ} = \emptyset$.
Bước (2): Vì điều kiện trong câu truy vấn là đơn điều kiện nên áp dụng Thuật toán 3.1 ta thực hiện câu truy vấn $select * from\ Nhanvien\ where\ TUOI =_{2(TUOI)} khả\ năng\ trẻ$. Theo Ví dụ 3.1 ta có kết quả sau.

Bảng 3.4. Kết quả thực hiện truy vấn trong Ví dụ 3.3 (i) chưa sử dụng lượng từ

STT	TENNV	NGHENGHIEP	TUOI	LUONG
1	An	Giáo viên	45	*rất cao*
3	Hà	Bác sĩ	*rất khả năng trẻ*	500
5	Nhân	Giáo viên	46	1500

Bước (3)-(12): Tiếp theo chúng ta đi đánh giá lượng từ *ít nhất 5* theo Thuật toán 3.3. Vì lượng từ $ít\ nhất\ 5 \in Q_{ABS}$ và đơn điệu tăng, ta có $f^A_{ít\ nhất\ 5}(\|r_{result}\|) = f^A_{ít\ nhất\ 5}(3) = 0$ nên kết quả của câu truy vấn $select * from\ Nhanvien\ whereít\ nhất\ 5(TUOI =_{2(LUONG)} khả\ năng\ trẻ)$ không có bộ nào.
Bước (13): Vậy $r_{resultQ} = \emptyset$.

(ii) Cho biết *hầu hết những cán bộ Tuổi khả năng trẻ và có Lương rất cao*, khi đó câu truy vấn SQL mờ: $select * from\ Nhanvien\ where\ \ Hầu\ hết\ (TUOI =_{2(TUOI)} khả\ năng\ trẻ\ and\ LUONG =_{2(LUONG)} rất\ cao)$.
Bước (1): $r_{resultQ} = \emptyset$.

Bước (2): Vì điều kiện trong câu truy vấn là đa điều kiện nên áp dụng Thuật toán 3.2 ta thực hiện câu truy vấn $select * from\ Nhanvien\ where\ (TUOI =_{2(TUOI)}$ *khả năng trẻ*) *and* $(LUONG =_{2(LUONG)}$ *rất cao*). Theo Ví dụ 3.2 ta có kết quả sau.

Bảng 3.5. Kết quả thực hiện truy vấn trong ví dụ 3.3 (ii) chưa sử dụng lượng từ

STT	TENNV	NGHENGHIEP	TUOI	LUONG
1	An	Giáo viên	45	*rất cao*
5	Nhân	Giáo viên	46	1500

Bước (3)-(12): Vì lượng từ *Hầu hết* $\in Q_{PRO}$ nên đi xây dựng các khoảng $S(\mathbf{1}), S(lớn), S(\boldsymbol{W}), S(nhỏ), S(\mathbf{0})$. Chọn $fm(lớn) = 0,35$, $fm(nhỏ) = 0,65$, $\mu(khả\ năng) = 0,25$, $\mu(ít) = 0,2$, $\mu(hơn) = 0,15$ và $\mu(rất) = 0,4$. Ta phân hoạch đoạn $[0,1]$ thành 5 khoảng tương tự mức 1 là:

$fm(rất\ lớn) = 0,35 \times 0,35 = 0,1225$. Vậy $S(\mathbf{1}) = (0,8775; 1]$.

$(fm(khả\ năng\ lớn) + fm(hơn\ lớn)) = (0,25 \times 0,35 + 0,15 \times 0,35) = 0,14$. Vậy $S(lớn) = (0,7375; 0,8775]$.

$(fm(ít\ nhỏ) + fm(ít\ lớn)) = (0,25 \times 0,65 + 0,25 \times 0,35) = 0,25$. Vậy $S(\boldsymbol{W}) = (0,4875; 0,8775]$.

$(fm(khả\ năng\ nhỏ) + fm(hơn\ nhỏ)) = (0,25 \times 0,65 + 0,15 \times 0,65) = 0,26$.

Vậy $S(nhỏ) = (0,2275; 0,4875]$ và $S(\mathbf{0}) = [0; 0,2275]$.

Vì $(\|r_{result}\|/\|r\|) = (2/9) = 0,222 \notin S(\mathbf{1})$ nên kết quả câu truy vấn

$$select * from\ Nhanvien\ where\ Hầu\ hết(TUOI =_{2(TUOI)} khả\ năng\ trẻ\ \ and\ LUONG =_{2(LUONG)} rất\ cao)$$

không có bộ nào thoả mãn.

Bước (13): Vậy $r_{resultQ} = \emptyset$.

4. KẾT LUẬN

Trong bài báo này, ngoài việc hệ thống một cách ngắn gọn một số kết quả trong [4] làm cơ sở cho việc nghiên cứu tiếp theo, chúng tôi đã nghiên cứu và đề xuất một phương pháp xử lý câu truy vấn SQL mờ dựa trên mô hình CSDL mờ theo cách tiếp cận ĐSGT. Các thuật toán và các ví dụ được trình bày rõ ràng, trực quan. Vấn đề lượng từ ngôn ngữ cũng được chúng tôi đề cập theo cách tiếp cận riêng và đưa lượng từ ngôn ngữ vào trong truy vấn SQL mờ làm cho vấn đề xử lý câu hỏi được mềm dẻo và gần gũi với thực tế. Cách xử lý truy vấn theo cách tiếp cận ĐSGT đơn giản và có nhiều ưu điểm so với các cách tiếp cận giải quyết trước đây như lý thuyết tập mờ, quan hệ tương tự và lý thuyết khả năng.

TÀI LIỆU THAM KHẢO

1. Nguyễn Cát Hồ, Trần Thái Sơn, Về khoảng cách giữa các giá trị của biến ngôn ngữ trong đại số gia tử, *Tạp chí Tin học và Điều khiển học* **11** (1) (1995) 10–20.

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

3. 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.
4. N. C. Ho, A model of relational databases with linguistic data of hedge algebrras based semantics, *Hội thảo quốc gia lần thứ ba về Nghiên cứu phát triển và ứng dụng CNTT và Truyền thông ICT.rda*, 2006.
5. Phương Minh Nam, Trần Thái Sơn, Về một cơ sở dữ liệu mờ và ứng dụng trong quản lý tội phạm, *Tạp chí Tin học và Điều khiển học* **22** (1) (2006) 25–36.
6. Nguyễn Công Hào, Mô hình cơ sở dữ liệu mờ theo cách tiếp cận đại số gia tử, *Kỷ yếu hội thảo quốc gia về các vấn đề chọn lọc công nghệ thông tin và truyền thông*, Hải Phòng, 2005 (285–293).
7. Nguyễn Công Hào, Truy vấn trong cơ sở dữ liệu mờ theo cách tiếp cận đại số gia tử, *Báo cáo toàn văn hội thảo khoa học một số vấn đề thời sự trong công nghệ thông tin và ứng dụng toán học*, Học viện kỹ thuật quân sự, 2006 (218–229).
8. Hồ Thuần, Hồ Cẩm Hà, Đại số quan hệ và quan điểm sử dụng Null value trên một mô hình cơ sở dữ liệu mờ, *Tạp chí Tin học và Điều khiển học* **17** (4) (2001) 1–10.
9. L. A. Zadeh, A computational approach to fuzzy quantifiers in natural languages, *Computers and Mathematics with Applications* **9** (2) (1983) 149–184.
10. Yoshikane Takahashi, Fuzzy database query language and their relational completeness theorem, *IEEE transactions on knowledge and data engineering* **5** (1) (1993) 122–125.
11. Hiroshi NAKAJIMA, Taiji SOGOH, Masaki ARAO, Fuzzy database query language and library-fuzzy extention to SQL, *1993 IEEE* (477–482).
12. P. Bosc, O. Pivert, SQLf: A relational database language for fuzzy query, *IEEE Transaction on Fuzzy Systems* **3** (1) (1995) (1–19).
13. Qi Yang, Chengwen Liu, Jing Wu, Clement Yu, Son Dao, Hiroshi Nakajima, Efficient processing of nested fuzzy SQL queries, *1995 IEEE* (131–138).
14. Patrick Bosc, Subqueries in SQLf, a fuzzy database query language, *1995 IEEE* (3636–3641).

Nhận bài ngày 13 - 10 - 2006
Nhận lại sau sửa ngày 7 - 12 - 2007