BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN VĂN PHONG

BIỂU DIỄN DỮ LIỆU MỜ BẰNG XML VÀ ỨNG DỤNG

Chuyên ngành: **KHOA HỌC MÁY TÍNH**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS Võ Trung Hùng**

Phản biện 1: PGS.TSKH. Trần Quốc Chiến

Phản biện 2: PGS.TS. Đoàn Văn Ban

Luận văn đã được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 04 tháng 03 năm 2012

Có thể tìm hiểu luận văn tại:
- Trung tâm Thông tin – Học liệu, Đại học Đà Nẵng.
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Theo hiểu biết của chúng tôi, hầu hết các nghiên cứu về cơ sở dữ liệu mờ chỉ xây dựng trên mô hình lý thuyết hoặc được cài đặt trên các mô hình quan hệ cổ điển mà chưa có một mô hình CSDL mờ thực sự được cài đặt trên máy tính. Do đó ít nhiều hạn chế việc ứng dụng các kết quả lý thuyết thu được. Vì vậy trong luận văn này, chúng tôi đề xuất một cách biểu diễn cơ sở dữ liệu mờ bằng XML, mô hình cơ sở dữ liệu mờ được biểu diễn là mô hình dựa trên lý thuyết về đại số gia tử.

2. Mục tiêu của đề tài

Mục đích nghiên cứu của đề tài là ứng dụng lý thuyết về mô hình cơ sở dữ liệu mờ, sử dụng ngôn ngữ XML để biểu diễn nhiều dạng dữ liệu mờ khác nhau, cho phép mờ hóa cơ sở dữ liệu sẵn có nhằm thu thập, lưu trữ và thực hiện các truy vấn trên cơ sở dữ liệu mờ ứng dụng trong việc khai phá dữ liệu nhằm đưa ra các dự báo trong tương lai.

3. Đối tượng nghiên cứu

Đối tượng mà đề tài nghiên cứu bao gồm việc tìm hiểu một số vấn đề nảy sinh trong quá trình quản lý thông tin nhân sự, nghiên cứu về đại số gia tử và mô hình cơ sở dữ liệu mờ dựa trên lý thuyết về đại số gia tử.

4. Phương pháp nghiên cứu

Đề tài thực hiện dựa trên nhiều phương pháp nghiên cứu khác nhau: khảo sát tình hình thực tế về các vấn đề về sử dụng những thông tin không đầy đủ, không chắc chắn trong thực tế, vấn đề về lưu trữ và xử lý những thông tin đó, tìm hiểu về cách xử lý thông tin nhân sự, nghiên cứu lý thuyết về cơ sở dữ liệu mờ dựa trên lý thuyết về đại số gia tử và ngôn ngữ XML.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Về mặt ý nghĩa khoa học và thực tiễn của đề tài là xây dựng những chức năng cho phép thu thập, lưu trữ những thông tin không chắc chắn, không đầy đủ; cho phép lưu trữ, xử lý và thực hiện truy vấn trên những thông tin đó; góp phần quan trọng trong lĩnh vực khai thác thông tin đặc biệt là những thông tin mờ. Kết quả này còn tiếp tục phát triển cho các tính toán và khai thác tri thức từ cơ sở dữ liệu mờ.

6. Bố cục của luận văn

Bố cục của luận văn gồm: Phần mở đầu. Chương 1, trình bày những khái niệm cơ bản. Chương 2, nghiên cứu ứng dụng logic mờ theo lý thuyết về đại số gia tử. Chương 3, xây dựng ứng dụng đưa ra các modul của bài toán. Kết luận và kiến nghị.

Chương 1. TỔNG QUAN

1.1. ĐẠI SỐ GIA TỬ

1.1.1. Một số khái niệm

1.1.2. Các tính chất của độ đo tính mờ trong ĐSGT

Mệnh đề 1.2. [5]

(1) $fm(hx) = \mu(h)fm(x)$, *với* $\forall x \in X$

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

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

(4) $\sum_{-q\le i\le p, i\neq 0} fm(h_i x) = fm(x)$, *với* $\forall x \in X$

(5) $\sum_{i=-q}^{-1} \mu(h_i) = \alpha$ và $\sum_{i=1}^{p} \mu(h_i) = \beta$, với $\alpha, \beta > 0$ và $\alpha + \beta = 1$

Định lý 1.2. [5][9] Cho $\underline{X} = (X, G, H, \preceq)$ là một ĐSGT tuyến tính. Ta có các phát biểu sau:

(1) Với $\forall x \in X$, $H(x)$ là tập sắp thứ tự tuyến tính.

(2) Nếu G là tập sắp thứ tự tuyến tính thì $H(G)$ cũng sắp thứ tự tuyến tính.

1.2. MÔ HÌNH BIỂU DIỄN CSDL MỜ THEO CÁCH TIẾP CẬN ĐSGT

Cho một ĐSGT tuyến tính đầy đủ $\mathcal{AX} = (X, G, C, H, \Phi, \Sigma, \leq)$, trong đó $Dom(\mathcal{X}) = X$ là miền các giá trị ngôn ngữ của thuộc tính ngôn ngữ $\mathcal{X}$ được sinh ra từ tập các phần tử sinh $G = \{0, c^-, W, c^+, 1\}$ bằng việc tác động các gia tử trong tập H, Σ và Φ là hai phép tính với ngữ nghĩa là cận trên đúng và cận dưới đúng của tập $H(x)$, tức là $\sum x = supremum\, H(x)$ and $\Phi x = infimum\, H(x)$, quan hệ $\preceq$ là quan hệ sắp thứ tự tuyến tính trên X cảm sinh từ ngữ nghĩa của ngôn ngữ [5].

1.2.1. Ngữ nghĩa dữ liệu dựa trên việc định lượng Đại số gia tử

1.2.1.1. Đặt vấn đề

1.2.1.2. Ngữ nghĩa dữ liệu dựa trên việc định lượng ĐSGT

Định nghĩa 1.5. [5] Cho $X_k = \{x \in X : |x| = k\}$, xét $P^k = \{I(x) : x \in X_k\}$ là một phân hoạch của [0, 1]. Gọi $\mathcal{V}$ là hàm định lượng ngữ nghĩa trên X.

(1) u bằng v theo mức k, được ký hiệu $u =_k v$, khi và chỉ khi $I(u)$ và $I(v)$ cùng chứa trong một khoảng mờ mức k. Có nghĩa là với $\forall u, v \in X, u =_k v \Leftrightarrow \exists \Delta^k \in P^k : I(u) \subseteq \Delta^k$ và $I(v) \subseteq \Delta^k$.

(2) u khác v theo mức k, được ký hiệu $u \neq_k v$, khi và chỉ khi $I(u)$ và $I(v)$ không cùng chứa trong một khoảng mờ mức k.

(3) u nhỏ hơn v theo mức k, được ký hiệu $u <_k v$, khi và chỉ khi $I(u)$ và $I(v)$ không cùng chứa trong một khoảng mờ mức k và $\mathcal{V}(u) < \mathcal{V}(v)$.

1.2.2. Phương pháp xử lý giá trị khoảng

Một cách tổng quát, nếu là giá trị a ta chuyển thành $[a, a]$, nếu giá trị là *khoảng* a ta chuyển thành $[a-\varepsilon, a+\varepsilon]$, với ε được xem là bán kính với tâm a. Nếu giá trị từ a đến b, thì được chuyển thành $[a, b]$. Do đó, quan hệ *Thunhapcanhan* có thể chuyển thành quan hệ sau [5]:

1.2.2.1. Chuyển các giá trị khoảng về đoạn con [0, 1] tương ứng

Gọi $Dom(A_i) = [min, max]$ là miền trị kinh điển của thuộc tính mờ A_i trong một quan hệ, trong đó *min*, *max* tương ứng là giá trị nhỏ nhất và giá trị lớn nhất của $Dom(A_i)$. Trước hết, ta sử dụng hàm f để chuyển đổi giá trị thuộc $Dom(A_i)$ thành giá trị thuộc [0, 1]. Tiếp theo, khoảng $[a, b]$ được biến đổi thành đoạn con [0, 1] tương ứng khi sử dụng hàm f, hay $[f(a), f(b)] \subseteq [0, 1]$.

1.2.2.2. Đối sánh các giá trị khoảng

Cho ĐSGT $\underline{X}=(X, G, H, \preceq)$ và một giá trị khoảng $[a, b]$. Để so sánh một giá trị $x \in X$ với $[a, b]$, trước hết chuyển $[a, b]$ về đoạn con của [0,1]. Vì tính mờ của x là một đoạn con của [0,1], do đó để so sánh $x \in X$ và đoạn con [0,1], chúng ta chỉ cần dựa vào phần giao của hai đoạn con của [0,1] tương ứng [5].

Với $x \in X$, ký hiệu $I(x) \subseteq [0,1]$ và $|I(x)| = fm(x)$, $[I_a, I_b] = [f(a), f(b)] \subseteq [0, 1]$ tương ứng với việc chuyển đổi giá trị khoảng $[a, b]$ về đoạn con của [0,1].

(1) Với mỗi $[I_a, I_b]$ nếu tồn tại $x \in X$ sao cho $[I_a, I_b] \subseteq I(x)$ thì $[a, b] =_{|x|} x$.

(2) Với mỗi $[I_a, I_b]$ sao cho $[I_a, I_b] \not\subset I(x) \forall x, x_1 \in X$ thì:

Khi đó với x và x_1, giả sử $x<x_1$ nếu $|[I_a, I_b] \cap I(x)| \geq |[I_a, I_b]|/\mathcal{L}$ thì $[a, b] =_{|x|} x$

Ngược lại nếu $|[I_a, I_b] \cap I(x)| \geq |[I_a, I_b]|/\mathcal{L}$ thì $[a, b] =_{|x_1|} x_1$.

(3) Với mỗi $[I_a, I_b]$ nếu tồn tại $x \in X$ sao cho $[I_a, I_b] \cap I(x) = \emptyset$ thì:

Nếu tồn tại $z \in X$ sao cho $[I_a, I_b] \subseteq I(z)$ và $I(x) \subseteq I(z)$ thì $[a, b] =_{|z|} x$.

1.2.3. Ngữ nghĩa dữ liệu dựa trên lân cận tôpô của ĐSGT

1.2.3.1. Độ tương tự mức k

Chúng ta luôn luôn giả thiết rằng mỗi tập H^- và H^+ chứa ít nhất 2 gia tử. Xét X_k là tập tất cả các phần tử độ dài k. Dựa trên các khoảng mờ mức k và các khoảng mờ mức k+1 chúng ta mô tả không hình thức 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 mức 2 trên nó là $I(h_p c^-) \leq I(h_{p-1} c^-) \leq \cdots \leq I(h_2 c^-) \leq I(h_1 c^-) \leq V_A(c^-) \leq I(h_{-1} c^-) \leq I(h_{-2} c^-) \leq \cdots \leq I(h_{-q+1} c^-) \leq I(h_{-q} c^-)$ khoảng

. Khi đó, ta xây dựng phân hoạch về độ tương tự mức 1 gồm các lớp tương đương sau: $S(0) = I(h_p c^-)$; $S(c^-) = I(c^-) \setminus [I(h_{-q} c^-) \cup I(h_p c^-)]$; $S(W) = I(h_{-q} c^-) \cup I(h_{-q} c^+)$; và một cách tương tự, $S(c^+) = I(c^+) \setminus [I(h_{-q} c^+) \cup I(h_p c^+)]$ và $S(1) = I(h_p c^+)$.

Ta thấy, trừ hai điểm đầu mút $V_A(0) = 0$ và $V_A(1) = 1$, các giá trị đại diện $V_A(c^-)$, $V_A(W)$ và $V_A(c^+)$ đều là điểm trong tương ứng của các lớp tương tự mức 1 $S(c^-)$, $S(W)$ 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, $I(h_i c^+) = [V_A(\Phi h_i c^+), V_A(\Sigma h_i c^+)]$ 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: $S(h_i c^+) = I(h_i c^+) \setminus [I(h_p h_i c^+) \cup I(h_{-q} h_i c^+)]$, $S(\Phi h_i c^+) = I(h_{-q} h_{i-1} c^+) \cup I(h_{-q} h_i c^+)$ và $S(\Phi h_i c^+) = I(h_p h_i c^+) \cup I(h_p h_i c^+)$, 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.

1.2.3.2. Lân cận mức k của khái niệm mờ

Định nghĩa 1.8. [5] Cho U là tập vũ trụ các thuộc tính, r là quan hệ xác định trên U, giả sử t_1 và t_2 là hai bộ dữ liệu thuộc quan hệ r. Ta ký hiệu $t_1[A_i] =_k t_2[A_i]$ và gọi chúng bằng nhau mức k, nếu một trong các điều kiện sau xảy ra:

(1) Nếu $t_1[A_i], t_2[A_i] \in D_{A_i}$ thì $t_1[A_i] = t_2[A_i]$;

(2) Nếu một trong hai giá trị $t_1[A_i], t_2[A_i]$ là khái niệm mờ, chẳng hạn đó là $t_1[A_i]$, thì ta phải có $t_2[A_i] \in \Omega_k(t_1[A_i])$;

(3) Nếu cả hai giá trị $t_1[A_i], t_2[A_i]$ là khái niệm mờ, thì $\Omega_k(t_1[A_i]) = \Omega_k(t_2[A_i])$.

Định nghĩa 1.9. [5] Cho *U* là tập vũ trụ các thuộc tính, *r* quan hệ xác định trên *U*, giả sử $t_1$và t_2 là hai bộ dữ liệu thuộc quan hệ *r*. Khi đó

(1)Ta viết $t_1[A_i] \leq_k t_2[A_i]$, nếu $t_1[A_i] =_k t_2[A_i]$ hoặc $\Omega_k(t_1[A_i]) < \Omega_k(t_2[A_i])$;

(2)Ta viết $t_1[A_i] <_k t_2[A_i]$, nếu $\Omega_k(t_1[A_i]) < \Omega_k(t_2[A_i])$;

(3)Ta viết $t_1[A_i] >_k t_2[A_i]$, nếu $\Omega_k(t_1[A_i]) > \Omega_k(t_2[A_i])$;

Sau đây là định lý khẳng định họ các khoảng $\Omega_k(x)$ là một phân hoạch của *Dom*(A_i) và giá trị định lượng của $x \in X$ luôn là điểm trong của lân cận mức *k* của *x*.

1.2.4. Phụ thuộc dữ liệu trong cơ sở dữ liệu mờ

1.2.4.1. Phụ thuộc hàm mờ

1.2.4.2. Phụ thuộc hàm mờ với lượng từ ngôn ngữ

a) Phụ thuộc hàm mờ với lượng từ ngôn ngữ

b) Phụ thuộc đơn điệu

c) Phụ thuộc đơn điệu trong CSDL kinh điển

1.2.4.3. Phụ thuộc đơn điệu trong CSDL mờ

a) Phụ thuộc đơn điệu tăng mức k

b) Phụ thuộc đơn điệu giảm mức k

1.3. NGÔN NGỮ ĐÁNH DẤU MỞ RỘNG XML

1.3.1. Document Prolog (phần mở đầu tài liệu)

1.3.2. Phần nội dung của tài liệu XML

1.3.2.1. Thẻ

Thẻ là các từ giữa các ký tự "<" và ">". Đặc tả XML quy định rất rõ về cách đặt tên thẻ: có thể bắt đầu bằng ký tự, gạch chân (_),

hoặc dấu hai chấm (:), các ký tự kế tiếp có thể là ký tự, ký số, gạch nối, dấu chấm, dấu hai chấm nhưng không được là khoảng trắng.

1.3.2.2. Thẻ mở (thẻ bắt đầu) và thẻ đóng (thẻ kết thúc)

Thẻ mở bắt đầu bằng ký tự "<" và kết thúc bằng ký tự ">"; thẻ đóng bắt đầu bằng ký tự "</" và kết thúc bằng ký tự ">". Các thẻ luôn đi cặp với nhau, sao cho mọi thẻ mở đều có một thẻ đóng tương ứng.

1.3.2.3. Phần tử

Phần tử là toàn bộ thông tin từ đầu của một thẻ mở đến cuối của một thẻ đóng

1.3.2.4. Phần tử rỗng

Phần tử rỗng là phần tử chỉ có duy nhất một thẻ. Đây là trường hợp các phần tử không kèm theo dữ liệu và có dạng *<tên thẻ/>*, ví dụ như *<img/>*, *
*, *
*.

1.3.2.5. Phần tử gốc

Phần tử gốc là phần tử bắt đầu một tài liệu XML.

1.3.2.6. Thuộc tính (Attribute)

1.3.3. Định nghĩa kiểu tư liệu (DTD)

1.3.3.1. Định nghĩa các phần tử

1.3.3.2. Khai báo phần tử với #PCDATA

1.3.3.3. Khai báo phần tử chứa nhiều phần tử con

1.3.3.4. Định nghĩa phần tử rỗng

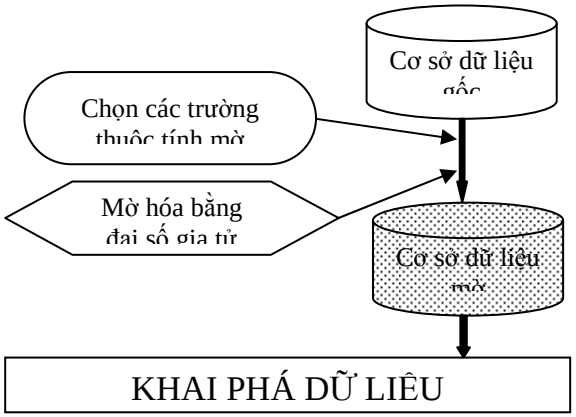
1.4. TỔNG KẾT CHƯƠNG

Các nội dung trong chương này tập trung giới thiệu về đại số gia tử, về mô hình cơ sở dữ liệu mờ dựa trên lý thuyết về đại số gia tử.

Chương 2. ĐỀ XUẤT ỨNG DỤNG

2.1. MÔ TẢ ỨNG DỤNG

2.2. MÔ HÌNH TỔNG QUÁT



Hình 2.1. *Mô hình tổng quát của hệ thống*

Bước 1: Chọn các trường thuộc tính trong cơ sở dữ liệu thông tin nhân sự.

Bước 2: Chọn các trường thuộc tính cần mờ hóa (thuộc tính mờ).

Bước 3: Mờ hóa cơ sở dữ liệu bằng ĐSGT.

Bước 4: Thực hiện các truy vấn trên cơ sở dữ liệu đã được mờ hóa.

2.3. ĐỀ XUẤT GIẢI PHÁP

Chúng tôi đề xuất giải pháp ứng dụng logic mờ trong hệ thống "*quản lý thông tin nhân sự*" dựa trên lý thuyết về đại số gia tử.

2.3.1. Mờ hóa cơ sở dữ liệu bằng đại số gia tử

Dựa trên cơ sở dữ liệu có sẵn chúng ta thực hiện mờ hóa dữ liệu dựa trên lý thuyết về đại số gia tử hay nói cách khác chúng ta cần xác định tập các phần tử sinh, tập gia tử và miền giá trị của nó, biểu diễn dữ liệu bằng tập các khoảng giá trị tương ứng. Trong đó mỗi thuộc tính trong cơ sở dữ liệu chuẩn XML cần phải xác định tập các phần tử sinh, tập gia tử và miền giá trị của nó, tính toán giá trị khoảng thông qua độ đo tính mờ và miền giá trị thuộc tính.

Sau khi tính được độ đo tính mờ, mỗi phần tử trong miền giá trị ngôn ngữ sẽ được biểu diễn thành các khoảng dữ liệu tương ứng. Dựa vào mục 1.2.3.1 ta xây dựng phân hoạch về độ tương tự mức 1 gồm các lớp tương đương sau: $S(0) = I(h_p c^-)$; $S(c^-) = I(c^-) \setminus [I(h_{-q}c^-) \cup I(h_p c^-)]$; $S(W) = I(h_{-q}c^-) \cup I(h_{-q}c^+)$; và một cách tương tự, $S(c^+) = I(c^+) \setminus [I(h_{-q}c^+) \cup I(h_p c^+)]$ và $S(1) = I(h_p c^+)$.

Tương tự, với *k*=2, ta có thể xây dựng phân hoạch các lớp tương tự mức 2. $S(h_i c^+) = I(h_i c^+) \setminus [I(h_p h_i c^+) \cup I(h_{-q} h_i c^+)]$, $S(\Phi h_i c^+) = I(h_{-q} h_{i-1} c^+) \cup I(h_{-q} h_i c^+)$ và $S(\Phi h_i c^+) = I(h_p h_i c^+) \cup I(h_p h_i c^+)$, với *i* sao cho $-q \le i \le p$ 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, dữ liệu được lưu trữ trong cơ sở dữ liệu gốc không chỉ có dạng dữ liệu rõ mà còn nhiều dạng dữ liệu khác. Ở đây chúng tôi đưa ra một số kiểu dữ liệu được ứng dụng trong luận văn này:

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

Kiểu 2: Giá trị rõ (tuổi bằng 49 hoặc tên là Nam)

Kiểu 3: Giá trị khoảng (tuổi nằm trong khoảng $25 \le t \le 27$)

Kiểu 4: Tập hữu hạn các giá trị rõ (tuổi là một trong số các số thuộc {29, 30, 31})

Kiểu 5: Giá trị không xác định (undefine).

Các dạng dữ liệu mờ này sẽ được chuyển về tập các khoảng khi thực hiện mờ hóa cơ sở dữ liệu gốc. Theo phương pháp sau:

Cho một ĐSGT tuyến tính đầy đủ, tập các giả tử H^- và H^+ có ít nhất hai phần tử. Khi đó, họ các khoảng $\{\Omega_k(x): x \in X\}$ được gọi là lân cận mức *k* của miền trị ngôn ngữ của thuộc tính A_i và là một phân hoạch của $Dom(A_i)$. Hơn nữa, mỗi giá trị *x* của A_i có duy nhất một lân cận mức *k*, $V_{A_i}(x)$ là điểm trong của $\Omega_k(x)$ với mọi $x \in X$. Khi đó các kiểu dữ liệu được biểu diễn lại như sau:

Kiểu 1: Mỗi giá trị *x* là dữ liệu mờ, khi đó họ các khoảng của *x* là $\Omega_k(x)$.

Kiểu 2 : Mỗi giá trị thực *x* là dữ liệu rõ, độ mờ của dữ liệu bằng 0, sẽ được biểu diễn bằng [*x*,*x*], tương ứng với mức mờ luôn luôn là ∞ nên còn gọi là khoảng mờ mức ∞ của a.

Vì vậy $\Omega_k(x) = [x, x]$. với mọi $1 \le k \le k^*$.

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 $\Omega_k([a,b]) = [a,b]$ với mọi $1 \le k \le k^*$.

Kiểu 4 : 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 2, $\Omega_k(P) = \{[a,a] \mid a \in P\}$ với mọi $1 \le k \le k^*$.

Kiểu 5 : Mỗi giá trị không được xác định (undefine, inapplicable) được biểu diễn bằng tập ϕ , xem như thông tin chính xác. Vì vậy $\Omega_k(inapplicable) = \{\phi\}$ với mọi $1 \le k \le k^*$

Cơ sở dữ liệu sau khi được mờ hóa bằng đại số gia tử sẽ được biểu diễn bằng ngôn ngữ XML để có thể lưu trữ và ứng dụng cho việc thực hiện các truy vấn mờ, khai phá tri thức và dự đoán sau này.

2.3.2. Biểu diễn dữ liệu mờ hóa bằng XML

2.3.2.1. Thẻ Attribute

Được dùng để xác định phần định nghĩa các thuộc tính của một cơ sở dữ liệu mờ: *<Attribute> … </Attribute>*

2.3.2.2. Thẻ Field

Được dùng để liệt kê các thuộc tính mờ của cơ sở dữ liệu. Thẻ Field nằm trong phạm vi của *thẻ Attribute*: *<Field>thuộc-tính</Field>*.

2.3.2.3. Thẻ Type

Được dùng để xác định kiểu dữ liệu của thuộc tính, kiểu dữ liệu có thể nhận là các kiểu dữ liệu nguyên thủy như: kiểu số nguyên, kiểu số thực, kiểu logic, kiểu chuỗi ký tự …. Thẻ *Type* nằm trong phạm vi của thẻ *Attribute*: *<Type>Kiểu-dữ-liệu</Type>*.

2.3.2.4. Thẻ D

Được dùng để xác định miền giá trị tham chiếu (qua giá trị được cho bởi thẻ <Min></Min> và <Max></Max>) cho các thuộc tính mờ. Thẻ *D* nằm trong phạm vi của thẻ *Attribute* và chỉ dùng cho các thuộc tính mờ:*<D><Min>…</Min></D>*.

***2.3.2.5. Thẻ* LDom**

Được dùng để xác định miền giá trị ngôn ngữ cho các thuộc tính mờ. Trong đó tập các phần tử sinh được liệt kê trong phạm vi của thẻ *<C></C>*, tập các gia tử được liệt kê trong phạm vi của thẻ *<H></H>*. Thẻ *LDom* nằm trong phạm vi của thẻ *Attribute* và chỉ dùng cho các thuộc tính mờ.

```
<LDom>
    <C>
          <Val Poss=fm(x)> x </Val>
          …..
    </C>
     <H>
          <Val Poss=fm(y) Type= H+/H-> y
</Val>
          …..
    </H>
</LDom>
```

Trong đó *fm(x), fm(y)* là độ đo mờ của biến ngôn ngữ *x, y. H+/ H-* *để xác định gia tử dương hay gia tử âm. Nếu Type=”H+” là gia tử dương, nếu Type=”H-” là gia tử âm. Trong phạm vi của thẻ <H></H> thì thứ tự của các gia tử được sắp xếp tăng dần theo* quan hệ cảm sinh ngữ nghĩa.

2.3.2.6. Thẻ Dist

Được dùng để xác định phạm vi của giá trị ngôn ngữ mờ


```
<Dist Type=”n”>
      {Phần khai báo giá trị ngôn ngữ mờ}
</Dist>
```

Trong đó *n* được dùng để xác định các kiểu dữ liệu thuộc 1 trong 5 kiểu dữ liệu mờ được nêu trong bài toán.

2.3.2.7. Thẻ Interval

Được dùng để xác định tập các giá trị khoảng của biến ngôn ngữ:

```
<Interval>
      <I Min=”x1” Max=”y1”></I>
      <I Min=”x2” Max=”y2”></I>
      …….
      <I Min=”xm” Max=”ym”></I>
</Interval >
```

Trong đó (x_i , y_i) là giá trị khoảng của biến ngôn ngữ.

2.3.3. Truy vấn trên dữ liệu mờ

2.3.3.1. Biểu diễn truy vấn

Để thực hiện các truy vấn mờ trên cơ sở dữ liệu mờ biểu diễn bằng XML thì trước tiên ta phải chuyển các truy vấn mờ về truy vấn rõ sau đó sử dụng phương pháp đối sánh mờ mức *k* được trình bày trong mục 1.2.3.2 để thực hiện truy vấn. Cấu trúc của lệnh truy vấn mờ dựa trên cơ sở các câu lệnh SQL nhưng ở đây chúng tôi quy định một số thẻ trong XML để mô tả câu lệnh truy vấn mờ.

Tương tự như trong CSDL quan hệ, dạng tổng quát của câu lệnh truy vấn SQL sử dụng truy vấn trong CSDL mờ được biểu diễn như sau:

```
SELECT [DISTINCT]<danh sách cột>
FROM <danh sách các bảng>
[WHERE <biểu thức điều kiện>]
```

Khi đó câu lệnh SQL sẽ được biểu diễn thông qua các thẻ của XML.

a) Thẻ select

Được dùng để liệt kê các trường được chọn để thực hiện truy vấn. Tên các trường được liệt kê qua thẻ *Field*.

```
<Select>
      <Field>Truong_1</Field>
      …
      <Field>Truong_n</Field>
</Select>
```

b) Thẻ From

Được dùng để chỉ các bảng được chọn. Tên các bảng được liệt kê thông qua thẻ *Table*.

```
<From>
      <Table>Bang_1</Table>
      ….
      <Table>Bang_n</Table>
</From>
```

c) Thẻ Expression

Được dùng để biểu diễn một biểu thức điều kiện.

- Biểu thức điều kiện mờ

```
<Expression Type=”Fuzzy”>
      <Field>Tên-Trường</Field>
      <Math              val=”Phép-toán-so-
sánh”></Math>
      <Val           type=”n”>giá-trị-đối-
sánh</Val>
</ Expression >
```

- Biểu thức điều kiện rõ

```
<Expression Type=”UnFuzzy”>
      <Field>Tên-Trường</Field>
      <Math              val=”Phép-toán-so-
sánh”></Math>
      <Val>giá-trị-đối-sánh</Val>
</ Expression >
```

d) Thẻ Where

Được dùng để liệt kê các biểu thức điều kiện. Các biểu thức điều kiện được kết hợp với nhau thông quả thẻ *Math*. Thẻ *Math* được dùng để liệt kê phép toán kết hợp giữa các biểu thức điều kiện và nó chỉ nhận 2 giá trị hoặc là “*And*” hoặc là “*Or*”.

```
<Where>
      <Expression Type=…>…</Expression>
      <Math>{And, Or}</Math>
      <Expression Type=…>…</Expression>
      …..
</Where>
```

2.3.3.2. Đánh giá truy vấn

a) *Thuật toán xác định giá trị chân lý của điều kiện mờ*

Thuật toán 2.2. Xác định giá trị chân lý của đa điều kiện mờ với phép toán θ

Vào: cho *r* là một quan hệ xác định trên vũ trụ các thuộc tính $U=\{A_1, A_2, \ldots, A_n\}$.

Điều kiện $A_i\,\theta\, fvalue_i\, \xi\, A_j\, \theta_1\, fvalue_j$

Ra: Với mọi $t \in r$ thỏa mãn điều kiện $((t[A_i]\theta\ fvalue_i\ \xi\ t[A_j]\theta_1\ fvalue_j)$

Phương pháp

```
(1) Begin
(2)    for each t∈r do
(3)       Begin
(4)          If     t[Ai] ∈ DAi          then
t[Ai] = Φk(ℑ(t[Ai]))
(5)          If     t[Aj] ∈ DAj          then
t[Aj] = Φk(ℑ(t[Aj]))
(6)       End
// Xây dựng các P^k_Ai và P^k_Aj dựa vào độ dài
các từ.
(7)    K = 1
(8)    While k ≤ p do
(9)       Begin
(10)      P^k_Ai = ∅; P^k_Aj = ∅;
(11)      For each t∈r do
(12)         begin
```



```
(13)          If     |t[Ai]| = k          then
P^k_Ai = P^k_Ai ∪ {I(t[Ai])}
(14)          If     |t[Aj]| = k          then
P^k_Aj = P^k_Aj ∪ {I(t[Aj])}
(15)       End
(16)     K = k +1
(17)   End
(18) For each t∈r do
(19)   Begin
// Trường hợp ξ là phép toán and
(20)   If        ((t[Ai]θ fvalue_i) = 1)     and
((t[Aj]θ1 fvalue_j) = 1)                     then
  ((t[Ai]θ fvalue_i) and (t[Aj]θ1 fvalue_j) = 1
// Trường hợp ξ là phép toán or
(21)   If        ((t[Ai]θ fvalue_i) = 1)     or
((t[Aj]θ1 fvalue_j) = 1)  then
  ((t[Ai]θ fvalue_i) or (t[Aj]θ1 fvalue_j) = 1
(22)   End
(23)End.
```

b) Phương pháp truy vấn dữ liệu mờ

Câu lệnh SQL trong CSDL mờ có thể được tổng quát hóa sau:

(1) : Xác định giá trị chân lý của các điều kiện mờ (Sử dụng thuật toán 2.1, 2.2, 2.3, 2.4) và liên kết các giá trị chân lý vừa xác định.

(2) : Chọn các bộ dữ liệu thỏa mãn bước (1).

Do đó, vấn đề quan trọng của câu lệnh SQL trong CSDL mờ chính là xác định giá trị chân lý của điều kiện mờ và liên kết các giá trị chân lý đó.

2.4. TỔNG KẾT CHƯƠNG

Trong chương này, luận văn đã tập trung nghiên cứu các vấn đề nảy sinh trong hệ thống "*quản lý thông tin nhân sự*". Qua đó luận văn đã đề xuất hướng giải quyết và đưa ra các mô hình cơ sở dữ liệu mờ dựa trên lý thuyết về đại số gia tử đã được phân tích để giải quyết các yêu cầu của hệ thống, đồng thời trình bày phương pháp sử dụng ngôn ngữ XML để biểu diễn mô hình cơ sở dữ liệu mờ đó. Từ đó, luận văn đã đưa ra cách truy vấn mờ trên cơ sở dữ liệu mờ trên.

Chương 3. XÂY DỰNG ỨNG DỤNG

3.1. CÔNG CỤ LỰA CHỌN

Để hoàn thành ứng dụng này thì chương trình demo được viết trên ngôn ngữ lập trình C# trong bộ Visual Studio 2005 và ngôn ngữ XML, chương trình chạy trên hệ điều hành Windows XP, Vista, Windows 7.

3.2. PHÁT TRIỂN CÁC MODUL

3.3.1. Modul biểu diễn dữ liệu mờ

Ở đây, luận văn sử dụng lớp *FuzzyField* để lưu trữ các thuộc tính trong cơ sở dữ liệu mờ theo lý thuyết về đại số gia tử, trên mỗi thuộc tính mờ sẽ có các đặc trưng như phạm vi tham chiếu, tập các phần tử sinh, tập các gia tử,...

Trong đó miền giá trị tham chiếu được xác định bởi thuộc tính *min, max*; tập các phần tử sinh được xác định qua thuộc tính *CE*; tập các gia tử được xác định bằng thuộc tính *HE*.

Những thuộc tính này sẽ ràng buộc trên kiểu dữ liệu được đưa vào cơ sở dữ liệu mờ, trong đó các thông tin được đưa vào phải thuộc 1 trong 5 kiểu dữ liệu được đề xuất trong mục 2.2.1.1 và được biểu diễn thông qua lớp *SQLField*. Trong đó kiểu dữ liệu được quy định bởi thuộc tính *dataType*, dữ liệu được lưu trữ bởi thuộc tính *value* và được chuyển thành các khoảng *interval* nếu là trường thuộc tính mờ.

Để mờ hóa dữ liệu nguồn theo lý thuyết đại số gia tử, ta sử dụng phương thức *ExtendData* trong lớp *DataTable*. Với đầu vào là danh sách các trường được lựa chọn từ cơ sở dữ liệu nguồn, tương ứng với mỗi kiểu dữ liệu được đề xuất trong luận văn, phương thức này sẽ chuyển thành dữ liệu khoảng dựa trên lý thuyết về đại số gia tử.

Cơ sở dữ liệu sau khi được mờ hóa sẽ được lưu trữ trong một file xml.

Tuy nhiên, để có thể xử lý được dữ liệu mờ lưu trữ trong file xml thì việc đọc cơ sở dữ liệu mờ từ file xml cũng rất quan trọng. Chức năng này được thực hiện thông qua lớp *ReadXML*. Lớp *ReadXML* cho phép đọc danh sách các trường cùng với những thông tin của chúng và đưa vào đối tượng *MyField* (đối tượng quản lý danh sách trường thuộc tính) dữ liệu trên mỗi trường sẽ được đọc và đưa vào đối tượng *DataTable*.

3.3.2. Modul biểu diễn truy vấn mờ

Để thực hiện các truy vấn trên cơ sở dữ liệu mờ thì câu truy vấn ở dạng SQL phải được chuyển đổi thành file xml hoặc chuỗi theo cấu trúc của file xml. Công việc này sẽ được thực hiện thông qua phương thức *ReadSQLXML* trong lớp *SQL*.

Phương thức này cho phép chuyển đổi một câu lệnh SQL thành một file truy vấn biểu diễn bằng XML để có thể thực hiện truy vấn dễ dàng trên cơ sở dữ liệu mờ.

Truy vấn mờ qua phương thức *ExcuteQuery* trong lớp *SQL* kết quả truy vấn sẽ được lưu trữ trong đối tượng *DataTable* mà phương thức trả về.

3.4. GIAO DIỆN CHƯƠNG TRÌNH

Dựa trên những phân tích và thiết kế về cơ sở dữ liệu mờ theo lý thuyết về đại số gia tử, chúng tôi đã xây dựng hệ thống *"quản lý thông tin nhân sự"* trong đó có thêm các chức năng ứng dụng logic mờ dựa trên lý thuyết về đại số gia tử.

3.5. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Dựa trên 5 dạng dữ liệu mờ cơ bản chương trình đã thực hiện việc mờ hóa cơ sở dữ liệu nguồn, cho phép biểu diễn nhiều dạng dữ liệu mờ khác nhau khi cập nhật. Việc lưu trữ, truy vấn thực hiện nhanh chóng và cho kết quả đúng.

Với kết quả này chúng ta có thể sử dụng cho việc mờ hóa dữ liệu đã có, lưu trữ dữ liệu mờ trong nhiều hệ thống ứng dụng khác: Hệ thống khai phá dữ liệu tri thức bằng các luật kết hợp mờ, Hệ thống hỗ trợ quyết định, …

3.6. TỔNG KẾT CHƯƠNG

Nhằm áp dụng cơ sở dữ liệu mờ đã nghiên cứu vào giải quyết bài toán "*quản lý thông tin nhân sự*". Trong chương ba đã ứng dụng các kết quả của chương hai để bổ sung thêm chức năng ứng dụng logic mờ trong hệ thống "*quản lý thông tin nhân sự*" dựa trên lý thuyết về đại số gia tử. Chức năng này cho phép thực hiện mờ hóa cơ sở dữ liệu đã có, thu thập, lưu trữ cơ sở dữ liệu mờ để làm dữ liệu nguồn cho các hệ thống khai phá dữ liệu để đưa ra các dự báo trong tương lai và đưa ra các truy vấn mờ trên cơ sở dữ liệu đã được mờ hóa. Ngoài ra, trong chương ba cũng đưa ra kết quả thử nghiệm và đánh giá hệ thống này.

KẾT LUẬN VÀ KIẾN NGHỊ

Với mục đích tìm ra một phương pháp biểu diễn cơ sở dữ liệu mờ cho phép mờ hóa cơ sở dữ liệu sẵn có để thu thập, lưu trữ và xử lý được những thông tin mờ trên máy tính làm nguồn dữ liệu ứng dụng trong các hệ thống khai phá dữ liệu để đưa ra các dự báo có tính chiến lược trong tương lai. Với cách tiếp cận dựa trên những lý thuyết đã có về cơ sở dữ liệu mờ về ngôn ngữ biểu diễn dữ liệu. Luận văn đã đề xuất một phương pháp mới để biểu diễn cơ sở dữ liệu mờ có nhiều kiểu dữ liệu khác nhau dựa trên cấu trúc định lượng của ĐSGT bằng ngôn ngữ XML. Mỗi cơ sở dữ liệu mờ được biểu diễn theo một cấu trúc chung bao gồm: phần khai báo, các thuộc tính và phần nội dung bằng các thẻ XML.

Những nội dung chính mà luận văn đã tập trung nghiên cứu và giải quyết: lý thuyết về ĐSGT, mô hình cơ sở mờ dữ liệu dựa trên lý thuyết về ĐSGT, ngôn ngữ XML. Dựa trên cơ sở lý thuyết đó, luận văn đã vận dụng logic mờ trong hệ thống "*quản lý thông tin nhân sự*" để giải quyết vấn đề về việc sử dụng cơ sở dữ liệu mờ để khai phá dữ liệu, đưa ra các dự báo trong tương lai.

Trong quá trình xây dựng hệ thống vẫn còn tồn tại một số vấn đề cần phải được phát triển: Xây dựng thuật toán xấp xỉ dữ liệu để biểu diễn được dữ liệu NULL, bổ sung một số thuật toán trong việc tìm kiếm các thẻ của XML nhanh hơn như: sử dụng Xpath và Xquery trong tìm kiếm và đối sánh dữ liệu, xây dựng một số modul cho phép thu thập nhiều dạng dữ liệu mờ, cải tiến phương pháp đối sánh mờ, vận dụng lý thuyết về đại số gia tử không thuần nhất trong các ứng dụng logic mờ,…

Với phương pháp này, bước đầu chúng tôi đã cài đặt thành công mô hình cơ sở dữ liệu mờ dựa trên lý thuyết về đại số gia tử, cho phép mờ hóa cơ sở dữ liệu đã có, thu thập thông tin, cập nhật thông tin và thực hiện một số tính toán, cũng như truy vấn mờ trên cơ sở dữ liệu đó.