Tạp chí Tin học và Điều khiển học, T. 19, S. 2 (2003), 17–27

MỘT SỐ MỞ RỘNG TỔNG KẾT DỮ LIỆU TRÊN CƠ SỞ DỮ LIỆU QUAN HỆ MỜ*

TRẦN THIÊN THÀNH

Khoa Tin học, Trường Đại học Sư phạm Quy Nhơn

Abstract. In this paper, we present some extensions of data summaries on fuzzy relational databases based on a pattern matching process of D. Dubois and H. Prade. An algorithm for discovering data rules based on hierarchical tree of template rules is given.

Tóm tắt. Trong bài báo này, chúng tôi trình bày một số mở rộng tính toán cho các luật tổng kết từ dữ liệu trên mô hình cơ sở dữ liệu quan hệ mờ. Các tính toán được xây dựng dựa trên đối sánh mẫu của D. Dubois và H. Prade. Bài báo cũng đưa ra thuật toán xây dựng các luật tổng kết dữ liệu dựa trên cây phân cấp của các luật mẫu.

1. ĐẶT VẤN ĐỀ

Việc phát hiện tri thức từ dữ liệu là một trong những hướng nghiên cứu đã tạo ra một cách nhìn mới về những dữ liệu được lưu trữ. Cùng với sự phát triển các mô hình cơ sở dữ liệu quan hệ mờ (cơ sở dữ liệu quan hệ với dữ liệu mờ) đã cho phép thu thập nhiều thông tin và qua đó hỗ trợ nhiều cho việc phát hiện các tri thức. Các tri thức phát hiện từ dữ liệu thường có dạng các luật dữ liệu bao gồm các yếu tố mờ như lượng từ mờ, tân từ mờ, các phép so sánh mờ, ... và các luật này có độ tin cậy là một giá trị trong đoạn [0,1]. Tiêu biểu cho hướng nghiên cứu này là các kết quả của Yager [9, 10], Kacprzyk [7], Cubero [3], Bosc [1], Dubois, Prade [6],....

Trong bài báo này, trên cơ sở đánh giá độ tin cậy của luật có dạng "*Q r are P*" của Dubois và Prade [6], chúng tôi xây dựng cách đánh giá độ tin cậy cho các luật có dạng "*Q* P_1 *r are* P_2" và "Q_1 P_1 *r* θ Q_2 P_2 *r*", trong đó Q, Q_1, Q_2 là các lượng từ mờ, r là một quan hệ mờ, P_1, P_2 là các tân từ mờ, θ là toán tử so sánh mờ. Cùng với cách đánh giá độ tin cậy, dựa vào thứ tự phân cấp của các tập mờ, chúng tôi đưa ra thuật toán nhằm xây dựng tập các luật tổng kết dữ liệu theo mẫu cho trước trên các dữ liệu có sẵn.

Bài báo được tổ chức như sau: phần 2 trình bầy những kiến thức cơ sở gồm: mô hình CSDLQH mờ dựa trên khả năng; đánh giá các tân từ mờ; biểu diễn các lượng từ ngôn ngữ bằng tập mờ; lực lượng tập mờ. Phần 3 trình bày cách đánh giá các luật tổng kết dữ liệu và một số tính chất liên quan. Phần 4 trình bầy thuật toán xây dựng các luật theo mẫu dựa vào thứ tự phân cấp của các tập mờ. Cuối cùng là kết luận và một số hướng nghiên cứu tiếp theo.

2. CÁC KIẾN THỨC CƠ SỞ

2.1. Mô hình cơ sở dữ liệu quan hệ mờ dựa trên khả năng

Bằng phân bố khả năng có thể biểu diễn dữ liệu của từng thuộc tính cho mỗi n-bộ. Giả sử A là một thuộc tính của lược đồ quan hệ, D là miền trị của A. Giá trị của một n-bộ t tại thuộc tính A được biểu diễn bởi phân bố khả năng chuẩn $\pi_{A(t)}$ trên miền trị mở rộng

* Công trình được hoàn thành với sự hỗtrợ kinh phí của Hội đòng khoa học tự nhiên.

$\widetilde{D} = D \cup \{e\}$, trong đó e là phần tử bổ sung vào mỗi miền trị, được sử dụng trong trường hợp thuộc tính A không áp dụng được cho bộ t (chi tiết xem trong [4]).

2.2. Tính toán trên các tân từ mờ

Cho X là một biến nhận giá trị trên miền D kết hợp với phân bố khả năng π_X, F là một tập mờ trên D. Độ tương thích của X với tập mờ F được đánh giá trên hai độ đo khả năng (Π) và cần thiết (N) được xác định bởi:

$$\Pi_X(F) = \sup_{u \in D} \min\Big(\mu_F(u), \pi_X(u)\Big) \tag{2.1}$$

$$N_X(F) = \inf_{u \in D} \max\Big(\mu_F(u), 1 - \pi_X(u)\Big). \tag{2.2}$$

Với θ là một phép so sánh mờ được xác định bởi hàm thuộc μ_θ, mệnh đề " $X\ \theta\ F$" được xem tương đương với mệnh đề " X *is* $F \circ \theta$", với $F \circ \theta$ là phép hợp thành của một giá trị mờ F với một toán tử so sánh mờ θ được xác định bởi: $\forall d \in D,\ \mu_{F\circ\theta}(d) = \sup\limits_{d' \in D} \min\Big(\mu_\theta(d, d'),\ \mu_F(d')\Big)$. Độ thỏa mệnh đề "$X\ \theta\ F$" được đánh giá như trong (2.1) và (2.2):

$$\Pi(X\theta F) = \Pi_X(F \circ \theta) = \sup_{u \in D} \min\Big(\mu_{F\circ\theta}(u), \pi_X(u)\Big) \tag{2.3}$$

$$N(X\theta F) = N_X(F \circ \theta) = \inf_{u \in D} \max\Big(\mu_{F\circ\theta}(u), 1 - \pi_X(u)\Big). \tag{2.4}$$

Dựa trên đánh giá của những tân từ nguyên tố, trong [4] đưa ra những công thức đánh giá cho tân từ kết hợp các tân từ nguyên tố bởi các phép toán logic not, and, or.

2.3. Các lượng từ mờ

Trong [19] Zadeh đề xuất một cách biểu diễn lượng từ ngôn ngữ theo cách tiếp cận của lý thuyết tập mờ, trong đó mỗi lượng từ Q được xem như một tập mờ trên tập cơ sở X và được xác định qua hàm thuộc $\mu_Q : X \to [0, 1]$, với X là tập số nguyên không âm hoặc đoạn [0, 1] tuỳ thuộc vào loại lượng từ.

Zadeh chia các lượng từ ngôn ngữ thành hai loại: *lượng từ tuyệt đối* (absolute quantifiers) và *lượng từ tỷ lệ* (proportional quantifiers). Lượng từ tuyệt đối dùng trong những mệnh đề với số lượng xác định như : "*khoảng 2*", "*nhiều hơn 5*",... Lượng từ tuyệt đối được biểu diễn bằng tập mờ trên tập cơ sở là tập các số nguyên không âm. Lượng từ tỷ lệ thể hiện những số lượng phụ thuộc vào số lượng tập các đối tượng mà nó thể hiện, như các lượng từ: "*hầu hết*", "*khoảng một nửa*", ... Với những lượng từ này biểu diễn bằng tập mờ trên miền cơ sở là đoạn [0,1].

Lượng từ Q gọi là *đơn điệu tăng* nếu với mọi $x_1 > x_2$ thì $\mu_Q(x_1) \geq \mu_Q(x_2)$. Chẳng hạn: "*at least 3*", "*almost all*", "*most*",...

Lượng từ Q gọi là *đơn điệu giảm* nếu với mọi $x_1 > x_2$ thì $\mu_Q(x_1) \leq \mu_Q(x_2)$. Chẳng hạn: "*at most 3*", "*few*", "*almost none*",...

Lượng từ Q gọi là lượng từ *unimodal* nếu tồn tại hai giá trị a, b với $a \leq b$ sao cho với mọi $x < a$ thì Q là lượng từ đơn điệu tăng; với $x > b$ thì Q đơn điệu giảm và $\mu_Q(x) = 1$ với mọi $x \in [a, b]$.

Nhận xét 2.1. Với mọi lượng từ unimodal Q bao giờ cũng tìm được hai lượng từ Q_a đơn điệu tăng và Q_d đơn điệu giảm sao cho $Q = Q_a \cap Q_d$.

Lượng từ phủ định (negation) của một lượng từ Q, ký hiệu $\overline{Q}$, được xác định bởi $\mu_{\overline{Q}}(x) = 1 - \mu_Q(x)\ \forall x$. Chẳng hạn *not many* là lượng từ phủ định của lượng từ *many*.

2.4. Lực lượng mờ

Có nhiều cách tiếp cận để định nghĩa lực lượng mờ, trong bài báo này chúng tôi dùng định nghĩa lực lượng mờ theo cách tiếp cận của Dubois và Prade [6].

Cho F là tập mờ trên tập hữu hạn $U = \{u_1, u_2, \cdots, u_n\}$. Đặt $k = |\ker(F)|$, với $\ker(F) = \{u \in U | \mu_F(u) = 1\}$. Lực lượng của tập mờ F, ký hiệu $|F|_f$ (hoặc $|F|$ nếu không gây nhầm lẫn) là một phân bố khả năng chuẩn $\pi_{|F|}$ trên đoạn $[0, n]$, được xác định như sau:

$$\pi(t) = 0 \quad \text{với} \quad 0 \leq t < k,$$

$$\pi(k) = 1,$$

với $j > k$ thì $\pi(j)$ là giá trị lớn thứ j trong danh sách các giá trị $\mu(u_1), \mu(u_2), \cdots, \mu(u_n)$.

Nhận xét 2.2. Nếu $F' \subseteq F$ thì với mọi $i \geq k$ ta có $\pi_{|F'|}(i) \leq \pi_{|F|}(i)$, với $k = |\ker(F)|$.

3. TỔNG KẾT DỮ LIỆU

Trong phần này xây dựng các công thức tính độ tin cậy của một luật tổng kết dữ liệu. Độ tin cậy được đánh giá trên hai độ đo khả năng và cần thiết.

3.1. Dạng *Q r are P*

Mệnh đề "*Q r are P*" có nghĩa là định lượng các bộ trong quan hệ r thỏa tân từ P ở mức độ nào đó tương thích với lượng từ Q. Chẳng hạn các mệnh đề :"*Có ít nhất 5 người trong CSDL có lương cao*" hay "*Hầu hết những người trong CSDL là trẻ*".

Độ tin cậy của mệnh đề "*Q r are P*" được đánh giá như độ tin cậy của mệnh đề "$|r_P|_f$ is Q", với r_P là tập các bộ của quan hệ r thỏa tân từ P ở mức độ khả năng (hoặc cần thiết). Theo công thức (2.1) và (2.2) trong trường hợp Q là lượng từ tuyệt đối ta có:

$$\Pi_{Q\ r\ are\ P} = \max_{k \leq i \leq n} \min\Big(\mu_Q(i),\ \pi_{|r_P|}(i)\Big) \tag{3.1}$$

$$N_{Q\ r\ are\ P} = \min_{k \leq i \leq n} \max\Big(\mu_Q(i),\ 1 - \pi_{|r_P|}(i)\Big), \tag{3.2}$$

với n là số bộ của quan hệ r, $k = |\ker(r_P)|$.

Nếu Q là lượng từ tỷ lệ, trong các công thức trên ta thay $\mu_Q(i)$ bởi $\mu_Q(i/n)$.

Định nghĩa 3.1. Cho P và P' là hai lượng từ mờ áp dụng trên lược đồ quan hệ r. Tân từ P' được gọi là yếu hơn tân từ P, ký hiệu $P' \subseteq P$ nếu với mọi quan hệ r của lược đồ r thỏa $\Pi(t|P') \leq \Pi(t|P)$ và $N(t|P') \leq N(t|P)$, với mọi bộ $t \in r$.

Dựa vào công thức (3.1) và (3.2) ta dễ dàng chứng minh được hai bổ đề sau:

Bổ đề 3.1. *a) Nếu Q là lượng từ đơn điệu tăng thì $N_{Q\ r\ are\ P} = \mu_Q(k)$.*
b) Nếu Q là lượng từ đơn điệu giảm thì $\Pi_{Q\ r\ are\ P} = \mu_Q(k)$, với $k = |ker(r_P)|$.

Bổ đề 3.2. *Với mọi lượng từ Q và tân từ P áp dụng trên quan hệ r, ta có*

$$\Pi_{\overline{Q}\ r\ are\ P} = 1 - N_{Q\ r\ are\ P} \quad \text{và} \quad N_{\overline{Q}\ r\ are\ P} = 1 - \Pi_{Q\ r\ are\ P}.$$

Định lý 3.1. *Nếu Q, Q' là các lượng từ đơn điệu tăng thỏa $Q' \subseteq Q$ và P, P' là các tân từ thỏa $P' \subseteq P$ thì với mọi quan hệ r, ta có $\Pi_{Q\ r\ are\ P} \geq \Pi_{Q'\ r\ are\ P'}$ và $N_{Q\ r\ are\ P} \geq N_{Q'\ r\ are\ P'}$.*

Chứng minh. Đặt $k = |\ker(r_P)|$, $k' = |\ker(r_{P'})|$. Vì $P' \subseteq P$ nên $k' \le k$. Từ Bổ đề 3.1. và do $Q' \subseteq Q$ là các lượng từ đơn điệu tăng nên dễ dàng suy ra $N_{Q\ r\ are\ P} \ge N_{Q'\ r\ are\ P'}$.

Ta có:

$$\Pi_{Q\ r\ are\ P} = \max_{k\le i\le n} \min\Big(\mu_Q(i), \pi_{|r_P|}(i)\Big) \quad \text{và} \quad \Pi_{Q'\ r\ are\ P'} = \max_{k'\le i\le n} \min\Big(\mu_{Q'}(i),\ \pi_{|r_{P'}|}(i)\Big).$$

Dễ thấy $\forall i,\ k \le i \le n$ thì $\min\Big(\mu_Q(i),\ \pi_{|r_P|}(i)\Big) \ge \min\Big(\mu_{Q'}(i),\ \pi_{|r_{P'}|}(i)\Big)$ và

$$\forall j,\ k' \le j < k \quad \text{thì} \quad \min\Big(\mu_Q(k),\ \pi_{|r_P|}(k)\Big) \ge \min\Big(\mu_{Q'}(j),\ \pi_{|r_{P'}|}(j)\Big).$$

Từ đó suy ra $\Pi_{Q\ r\ are\ P} \ge \Pi_{Q'\ r\ are\ P'}$. ∎

Tương tự, ta có định lý tương ứng cho các lượng từ đơn điệu giảm.

Định lý 3.2. *Nếu Q, Q' là các lượng từ đơn điệu giảm thỏa $Q \subseteq Q'$ và P, P' là các tân từ thỏa $P' \subseteq P$ thì với mọi quan hệ r, ta có:*

$$\Pi_{Q\ r\ are\ P} \le \Pi_{Q'\ r\ are\ P'} \quad \textit{và} \quad N_{Q\ r\ are\ P} \le N_{Q'\ r\ are\ P'}.$$

3.2. Dạng *Q P_1 r are P_2*

Mệnh đề "*Q* P_1 *r* are P_2" có nghĩa là định lượng các bộ trong quan hệ r thỏa tân từ P_1 cũng thỏa tân từ P_2 tương thích với lượng từ Q.

Trường hợp Q là lượng từ tuyệt đối thì ta có sự tương đương về mặt ngữ nghĩa của hai mệnh đề "*Q* P_1 *r are* P_2" và "*Q r are* P_1 *and* P_2". Do đó hoàn toàn có thể đánh giá theo công thức (3.1) và (3.2). Chẳng hạn mệnh đề "*Có ít nhất 5 người tuổi cao trong cơ sở dữ liệu cũng có lương cao*" tương đương với mệnh đề "*Có ít nhất 5 người trong cơ sở dữ liệu có tuổi cao và lương cao*".

Trường hợp Q là lượng từ tỷ lệ, ta có thể xem mệnh đề "*Q* P_1 *r are* P_2" tương đương với mệnh đề "*Q* r_{P_1} *are* P_2", với r_{P_1} là những bộ của quan hệ r thỏa tân từ P_1.

Ký hiệu $k_1 = |\ker(r_{P_1})|$, độ đo khả năng thỏa mệnh đề "*Q* P_1 *r are* P_2" được xây dựng qua các bước như sau:

Bước 1. Với mỗi i trong khoảng các giá trị có thể là lực lượng của quan hệ r_{P_1}, $k_1 \le i \le n$, gọi r^i là quan hệ được chọn từ r_{P_1} gồm i bộ có độ thỏa tân từ P_1 cao nhất. Khi đó độ tin cậy của mệnh đề "*Q* r_{P_1} *are* P_2" chính là độ tin cậy của mệnh đề "*Q* r^i *are* P_1 *and* P_2" được đánh giá theo công thức (3.1)

$$\Pi_{Q\ r^i\ are\ P_1\ and\ P_2} = \max_{k_2\le j\le i} \min\Big(\mu_Q(j/i),\ \pi_{|r^i_{P_1\wedge P_2}|}(j)\Big), \quad \text{với} \quad k_2 = |\ker(r_{P_1\wedge P_2})|.$$

Trong khi đó khả năng để quan hệ r_{P_1} có i bộ là $\pi_{|r_{P_1}|}(i)$ nên độ đo khả năng thỏa mệnh đề "*Q* r_{P_1} *are* P_2'' trong trường hợp r_{P_1} có đúng i bộ là:

$$\Pi_{Q\ r_{P_1}\ are\ P_2}(i) = \min\left(\max_{k_2\le j\le i}\Big\{\min\Big(\mu_Q(j/i),\ \pi_{|r^i_{P_1\wedge P_2}|}(j)\Big)\Big\}, \pi_{|r_{P_1}|}(i)\right).$$

Bước 2. Khả năng thỏa mệnh đề "*Q* P_1 *r are* P_2" được đánh giá trong trường hợp thuận lợi nhất, nên:

$$\begin{aligned}\Pi_{Q\ P_1\ r\ are\ P_2} &= \max_{k_1\le i\le n}\Big\{\Pi_{Q\ r_{P_1}\ are\ P_2}(i)\Big\}\\ &= \max_{k_1\le i\le n}\Big\{\min\Big[\max_{k_2\le j\le i}\min\Big(\mu_Q(j/i),\ \pi_{|r_{P_1\wedge P_2}|}(j)\Big),\ \pi_{|r_{P_1}|}(i)\Big]\Big\}. \qquad (3.3)\end{aligned}$$

Từ độ đo khả năng ta dễ dàng suy ra độ đo cần thiết là:

$$N_{Q\ P_1\ r\ are\ P_2} = \min_{k_1 \le i \le n}\Big\{\max\Big(\min_{k_2 \le j \le i}\max\Big(\mu_Q(j/i), 1-\pi_{|r^i_{P_1\wedge P_2}|}(j)\Big), 1-\pi_{|r_{P_1}|}(i)\Big)\Big\}. \tag{3.4}$$

Một số kết quả sau thể hiện thứ tự của các độ đo tương ứng với thứ tự các lượng từ và tân từ mờ.

Định lý 3.3. *Nếu Q, Q' là các lượng từ tuyệt đối, đơn điệu tăng thỏa $Q' \subseteq Q$ và P_1, P_1', P_2, P_2' là các tân từ thỏa $P_1' \subseteq P_1$, $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có:*

$$\Pi_{Q\ P_1\ r\ are\ P_2} \ge \Pi_{Q'\ P_1'\ r\ are\ P_2'} \quad và \quad N_{Q\ P_1\ r\ are\ P_2} \ge N_{Q'\ P_1'\ r\ are\ P_2'}.$$

Chứng minh. Vì Q, Q' là các lượng từ tuyệt đối nên $\Pi_{Q\ P_1\ r\ are\ P_2} = \Pi_{Q\ r\ are\ P_1\ and\ P_2}$, và $\Pi_{Q'\ P_1'\ r\ are\ P_2'} = \Pi_{Q'\ r\ are\ P_1'\ and\ P_2'}$.

Từ $P_1' \subseteq P_1$ và $P_2' \subseteq P_2$ suy ra $P_1'\ and\ P_2' \subseteq P_1\ and\ P_2$, theo kết quả Định lý 3.1. dễ dàng suy ra $\Pi_{Q\ P_1\ r\ are\ P_2} \ge \Pi_{Q'\ P_1'\ r\ are\ P_2'}$.

Hoàn toàn tương tự ta cũng chứng minh được $N_{Q\ P_1\ r\ are\ P_2} \ge N_{Q'\ P_1'\ r\ are\ P_2'}$. ■

Kết quả tương tự cho các lượng từ tuyệt đối, đơn điệu giảm thể hiện qua định lý sau.

Định lý 3.4. *Nếu Q, Q' là các lượng từ tuyệt đối, đơn điệu giảm thỏa $Q \subseteq Q'$ và P_1, P_1', P_2, P_2' là các tân từ thỏa $P_1' \subseteq P_1$, $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có*

$$\Pi_{Q\ P_1\ r\ are\ P_2} \le \Pi_{Q'\ P_1'\ r\ are\ P_2'} \quad và \quad N_{Q\ P_1\ r\ are\ P_2} \le N_{Q'\ P_1'\ r\ are\ P_2'}.$$

Với lượng từ tỷ lệ, các kết quả có một số thay đổi, cụ thể như sau:

Định lý 3.5. *Nếu Q, Q' là các lượng từ tỷ lệ, đơn điệu tăng thỏa $Q' \subseteq Q$ và P_2, P_2' là các tân từ thỏa $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có*

$$\Pi_{Q\ P_1\ r\ are\ P_2} \ge \Pi_{Q'\ P_1\ r\ are\ P_2'} \quad và \quad N_{Q\ P_1\ r\ are\ P_2} \ge N_{Q'\ P_1\ r\ are\ P_2'}.$$

Chứng minh. Từ công thức (3.3) ta có:

$$\Pi_{Q\ P_1\ r\ are\ P_2} = \max_{k_1 \le i \le n} \Pi_{Q\ r_{P_1}\ are\ P_2}(i) \text{ với } \Pi_{Q\ r_{P_1}\ are\ P_2}(i) = \min\Big(\Pi_{Q\ r^i\ are\ P_1\ and\ P_2}, \pi_{|r_{P_1}|}(i)\Big)$$

và

$$\Pi_{Q'\ P_1\ r\ are\ P_2'} = \max_{k_1 \le i \le n} \Pi_{Q'\ r_{P_1}\ are\ P_2'}(i) \text{ với } \Pi_{Q'\ r_{P_1}\ are\ P_2'}(i) = \min\Big(\Pi_{Q'\ r^i\ are\ P_1\ and\ P_2'}, \pi_{|r_{P_1}|}(i)\Big).$$

Theo Định lý 3.1. ta có $\Pi_{Q\ r^i\ are\ P_1\ and\ P_2} \ge \Pi_{Q'\ r^i\ are\ P_1\ and\ P_2'}$, $\forall i, k_1 \le i \le n$ nên

$$\Pi_{Q\ r_{P_1}\ are\ P_2}(i) \ge \Pi_{Q'\ r_{P_1}\ are\ P_2'}(i).$$

Do đó $\Pi_{Q\ P_1\ r\ are\ P_2} \ge \Pi_{Q'\ P_1\ r\ are\ P_2'}$.

Tương tự ta chứng minh được $N_{Q\ P_1\ r\ are\ P_2} \ge N_{Q'\ P_1\ r\ are\ P_2'}$. ■

Tương tự ta có định lý sau cho các lượng từ đơn điệu giảm.

Định lý 3.6. *Nếu Q , Q' là các lượng từ tỷ lệ, đơn điệu giảm thỏa $Q \subseteq Q'$ và P_2, P_2' là các tân từ thỏa $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có:*

$$\Pi_{Q\ P_1\ r\ are\ P_2} \le \Pi_{Q'\ P_1\ r\ are\ P_2'} \quad và \quad N_{Q\ P_1\ r\ are\ P_2} \le N_{Q'\ P_1\ r\ are\ P_2'}.$$

3.3. Dạng $Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r$.

Mệnh đề "$Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r$" có ý nghĩa là định lượng các bộ trong quan hệ r thỏa tân từ P_1 có quan hệ θ với các bộ trong quan hệ r thỏa tân từ P_2 ở mức độ của lượng từ Q_2 là tương thích với lượng từ Q_1. Chẳng hạn "*Hầu hết những người lớn tuổi trong CSDL có lương cao hơn nhiều người trẻ*".

Độ tin cậy của mệnh đề dạng này được đánh giá qua các bước như sau:

Bước 1. Với mỗi bộ $t_i \in r$, ta tính độ tin cậy của mệnh đề định lượng số bộ của quan hệ r thỏa tân từ P_2 có quan hệ θ với bộ t_i, $P_i =$ " $t_i\ \theta\ Q_2\ P_2\ r$" được biểu diễn tương đương với mệnh đề "$Q_2\ P_2\ r\ are\ t_i \circ \theta$". Do đó độ tin cậy của mệnh đề P_i được đánh giá theo công thức (3.3) và (3.4) :

$$\Pi_{P_i} = \max_{k_2 \le j \le n} \min\Big(\Pi_{Q_2\ r_{P_2}\ are\ t_i \circ \theta}(j), \pi_{|r_{P_2}|}(j)\Big) \tag{3.5}$$

$$N_{P_i} = \min_{k_2 \le j \le n} \max\Big(N_{Q_2\ r_{P_2}\ are\ t_i \circ \theta}(j), 1 - \pi_{|r_{P_2}|}(j)\Big) \tag{3.6}$$

với $k_2 = |\ker(r_{P_2})|$, $\Pi_{Q_2\ r_{P_2}\ are\ t_i \circ \theta}(j)$, $N_{Q_2\ r_{P_2}\ are\ t_i \circ \theta}(j)$ tương ứng là độ đo khả năng và cần thiết của mệnh đề "$Q_2\ r_{P_2}\ are\ t_i \circ \theta$" trong trường hợp r_{P_2} có đúng j bộ của r có độ thỏa tân từ P_2 cao nhất.

Bước 2. Ta xem "$\theta\ Q_2\ P_2\ r$" là một tân từ mờ trên các bộ của quan hệ r , ký hiệu tân từ này là P. Khi đó mệnh đề "$Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r$" được đưa về dạng tương đương "$Q_1\ P_1\ r\ are\ P$". Do đó độ tin cậy được đánh giá :

$$\Pi_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} = \max_{k_1 \le i \le n} \min\Big(\Pi_{Q_1\ r_{P_1}\ are\ P}(i), \pi_{|r_{P_1}|}(i)\Big) \tag{3.7}$$

$$N_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} = \min_{k_1 \le i \le n} \max\Big(N_{Q_1\ r_{P_1}\ are\ P}(i), 1 - \pi_{|r_{P_1}|}(i)\Big) \tag{3.8}$$

với $k_1 = |\ker(r_{P_1})|$, $\Pi_{Q_1\ r_{P_1}\ are\ P}(i)$, $N_{Q_1\ r_{P_1}\ are\ P}(i)$ tương ứng là độ tin cậy khả năng và cần thiết của mệnh đề "$Q_1\ r_{P_1}\ are\ P$" trong trường hợp r_{P_1} có đúng i bộ được chọn trong r có độ thỏa tân từ P_1 cao nhất.

Định lý 3.7. *Nếu Q_1, Q_1', Q_2, Q_2' là các lượng từ tuyệt đối, đơn điệu tăng thỏa $Q_1' \subseteq Q_1$, $Q_2' \subseteq Q_2$ và P_1, P_1', P_2, P_2' là các tân từ thỏa $P_1' \subseteq P_1$, $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có:*

$$\Pi_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \ge \Pi_{Q_1'\ P_1'\ r\ \theta\ Q_2'\ P_2'\ r} \quad \text{và} \quad N_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \ge N_{Q_1'\ P_1'\ r\ \theta\ Q_2'\ P_2'\ r}.$$

Chứng minh. Với mọi $t_i \in r$, gọi P_i là mệnh đề "$Q_2\ P_2\ r\ are\ t_i \circ \theta$", P_i' là mệnh đề " $Q_2'\ P_2'\ r\ are\ t_i \circ \theta$". Vì $P_2' \subseteq P_2$ nên theo định lý 3.3. ta có $\Pi_{P_i} \ge \Pi_{P_i'}$ và $N_{P_i} \ge N_{P_i'}$.

Do đó nếu gọi P là tân từ "$\theta\ Q_2\ P_2\ r$" và P' là tân từ "$\theta\ Q_2'\ P_2'\ r$" thì ta có $P' \subseteq P$.

Theo Định lý 3.3. suy ra

$$\Pi_{Q_1\ P_1\ r\ are\ P} \ge \Pi_{Q_1'\ P_1'\ r\ are\ P'} \quad \text{và} \quad N_{Q_1\ P_1\ r\ are\ P} \ge N_{Q_1'\ P_1'\ r\ are\ P'}. \qquad \blacksquare$$

Định lý sau cũng đúng cho các lượng từ tuyệt đối, đơn điệu giảm.

Định lý 3.8. *Nếu Q_1, Q_1', Q_2, Q_2' là các lượng từ tuyệt đối, đơn điệu giảm thỏa $Q_1 \subseteq Q_1'$, $Q_2 \subseteq Q_2'$ và P_1, P_1', P_2, P_2' là các tân từ thỏa $P_1' \subseteq P_1$, $P_2' \subseteq P_2$ thì với mọi quan hệ r ta có:*

$$\Pi_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \le \Pi_{Q_1'\ P_1'\ r\ \theta\ Q_2'\ P_2'\ r} \quad \text{và} \quad N_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \le N_{Q_1'\ P_1'\ r\ \theta\ Q_2'\ P_2'\ r}$$

Với các lượng từ tỷ lệ ta có các kết quả sau:

Định lý 3.9. *Nếu* Q_1, Q_1', Q_2, Q_2' *là các lượng từ tỷ lệ, đơn điệu tăng thỏa* $Q_1' \subseteq Q_1$, $Q_2' \subseteq Q_2$ *thì với mọi quan hệ* r *ta có:*

$$\Pi_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \geq \Pi_{Q_1'\ P_1\ r\ \theta\ Q_2'\ P_2\ r} \quad \text{và} \quad N_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \geq N_{Q_1'\ P_1\ r\ \theta\ Q_2'\ P_2\ r}.$$

Chứng minh. Dùng kết quả của Định lý 3.5 với kỹ thuật chứng minh tương tự như chứng minh của Định lý 3.7 ta dễ dàng chứng minh được định lý này. ■

Định lý 3.10. *Nếu* Q_1, Q_1', Q_2, Q_2' *là các lượng từ tỷ lệ, đơn điệu giảm thỏa* $Q_1 \subseteq Q_1'$, $Q_2 \subseteq Q_2'$ *thì với mọi quan hệ* r *ta có:*

$$\Pi_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \leq \Pi_{Q_1'\ P_1\ r\ \theta\ Q_2'\ P_2\ r} \quad \text{và} \quad N_{Q_1\ P_1\ r\ \theta\ Q_2\ P_2\ r} \leq N_{Q_1'\ P_1\ r\ \theta\ Q_2'\ P_2\ r}.$$

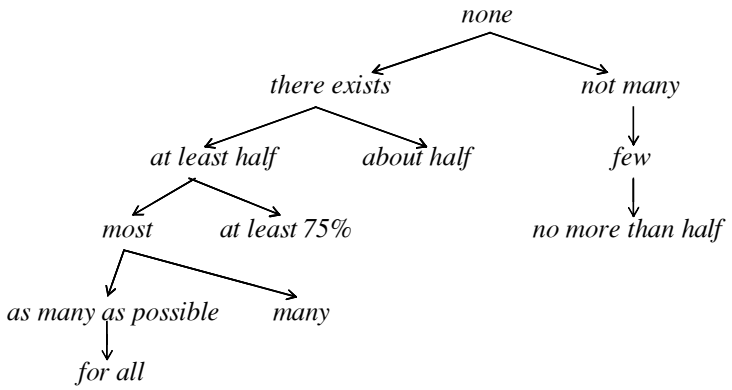
4. XÂY DỰNG CÁC LUẬT TỪ DỮ LIỆU

Một trong những nội dung quan trọng đặt ra cho việc phát hiện các luật từ dữ liệu là cần có những thuật toán tự động xây dựng các luật trên những dữ liệu cụ thể sao cho độ tin cậy vượt một ngưỡng cho trước nào đó. Tuy nhiên cho đến nay những thuật toán như vậy đều cần thiết phải có sự hỗ trợ một phần của con người. Trong phần tiếp theo chúng tôi trình bày một cách tiếp cận cho việc tổng kết dữ liệu tự động dựa theo mẫu và thứ tự phân cấp các tập mờ trong cùng miền trị.

4.1. Phân cấp các tập mờ

Cho là một tập hữu hạn các tập mờ trên miền D, khi đó cùng với quan hệ $\subseteq$ của các tập mờ tạo thành một thứ tự phân cấp. Với bất kỳ tập các tập mờ trên miền trị D ta bổ sung một tập mờ đặc biệt *none* được xác định $\mu_{none}(x) = 1, \forall x \in D$. Dễ thấy với mọi $F \in$ ta có $F \subseteq$ *none*. Tập mờ *none* được xem là gốc của cây thứ tự phân cấp các tập mờ.

Ví dụ 4.1. Giả sử là tập các lượng từ tỷ lệ = { *none, for all, there exists, most, at least half, as many as possible, many, at least 75%, about half, no more than half, not many, few*}. Cây thứ tự phân cấp các tập mờ trong như hình vẽ 1.



Hình 1. Cây phân cấp các lượng từ tỷ lệ

4.2. Luật mẫu và cây phân cấp

Từ các dạng luật được xem xét trong các mục 3.1, 3.2, 3.3, với một quan hệ r cụ thể, ta xem xét các luật mẫu với các tân từ nguyên tố có các dạng sau:

Dạng 1. *"Q A is F"*

Dạng 2. *"Q A is F also B is G"*

Dạng 3. "Q_1 *A is F* θ Q_2 *B is G*"

với Q, Q_1, Q_2 là lượng từ mờ, A, B là các thuộc tính của lược đồ quan hệ R, r là một quan hệ trên lược đồ R, F, G là các tập mờ tương ứng trên miền trị của thuộc tính A và B.

Phần này chỉ trình bày các nội dung liên quan đến luật mẫu dạng 1, các dạng còn lại có kết quả tương tự.

Xét luật mẫu có dạng "Q *A is F*", trong đó Q là một lượng từ nhận các giá trị trong một tập các lượng từ cho trước, A là một thuộc tính của lược đồ quan hệ r, F là một tập con mờ nhận giá trị trong một tập các tập mờ trên miền trị của thuộc tính A.

Từ những kết quả trong phần 3, ta có kết quả sau thể hiện thứ tự phân cấp của các luật:

Hệ quả 4.1. *Nếu Q, Q' là những lượng từ đơn điệu tăng thỏa $Q' \subseteq Q$ và $F' \subseteq F$ thì* $\Pi_{Q\ A\ is\ F} \geq \Pi_{Q'\ A\ is\ F'}$ *và* $N_{Q\ A\ is\ F} \geq N_{Q'\ A\ is\ F'}$.

Chứng minh. Với mỗi bộ $t \in r$, ký hiệu giá trị của bộ t tại thuộc tính A là phân bố khả năng $\pi_{A(t)}$.

Từ độ đo khả năng về tương thích của phân bố khả năng $\pi_{A(t)}$ với tập mờ F, F'

$\Pi(\pi_{A(t)}|F) = \sup\limits_{u \in D} \min\Big(\mu_F(u), \pi_{A(t)}(u)\Big)$ và $\Pi(\pi_{A(t)}|F') = \sup\limits_{u \in D} \min\Big(\mu_{F'}(u), \pi_{A(t)}(u)\Big)$

Do $F' \subseteq F$ nên ta có $\Pi(\pi_{A(t)}|F) \geq \Pi(\pi_{A(t)}|F')$.

Tương tự, dễ dàng kiểm chứng $N(\pi_{A(t)}|F) \geq N(\pi_{A(t)}|F')$.

Nếu xem P là tân từ "*A is F*" và P' là tân từ "*A is F'*" thì ta có thứ tự $P' \subseteq P$.

Theo Định lý 3.1. ta có $\Pi_{Q\ A\ is\ F} \geq \Pi_{Q'\ A\ is\ F'}$ và $N_{Q\ A\ is\ F} \geq N_{Q'\ A\ is\ F'}$. ■

Từ hệ quả trên, khi Q và F lần lượt nhận các giá trị tương ứng trong $_a$ (tập các lượng từ tăng) và với thứ tự phân cấp cho trước. Khi đó các cặp (Q, F) tạo ra một cây thứ tự phân cấp theo độ đo khả năng và cần thiết.

Ví dụ 4.2. Với luật mẫu Q *Age is* F, với $Q \in$ $_a$ (tập các lượng từ tăng trong Ví dụ 4.1), $F \in$ = {*none, young, old, middle, very young, very old, about 20, about 40, not young*}. Một phần cấu trúc cây phân cấp của luật mẫu có dạng như Hình 2.

Với lượng từ đơn điệu giảm ta cũng có kết quả tương tự dựa vào kết quả của Định lý 3.2. Từ đó ta có thuật toán xây dựng tập các luật cho luật mẫu dạng 1 ứng với tập lượng từ đơn điệu (tăng hoặc giảm).

4.3. Thuật toán xây dựng luật từ dữ liệu

Thuật toán 4.1. Xây dựng tập các luật từ luật mẫu dạng 1 cho tập lượng từ đơn điệu

Input : r là một quan hệ mờ

Luật mẫu RL = "Q *A is* F"

(, ⊆) là tập các lượng từ mờ đơn điệu với thứ tự phân cấp

(, ⊆) là tập các tập mờ trên thuộc tính A với thứ tự phân cấp

Ngưỡng xác định độ tin cậy α, β

Output: Tập các cặp (Q, F) thỏa $\Pi_{Q\ A\ is\ F}(r) \geq \alpha$ và $N_{Q\ A\ is\ F}(r) \geq \beta$.

Format : Rules(RL, , , α, β)

Method:

$H := (none, none)$;

$CS := \{H\}$;

$SS := \emptyset$;

While $CS \neq \emptyset$ **do**

$NextCS := \emptyset$;
For each H in CS **do**
If Sat(H, α, β) **then**
$SS := SS \cup \{H\}$;
For each $Hnext$ in Child(H) **do**
$NextCS := NextCS \cup \{Hnext\}$;
EndFor
Endif
EndFor
$CS := NextCS$;
EndWhile
Return SS;

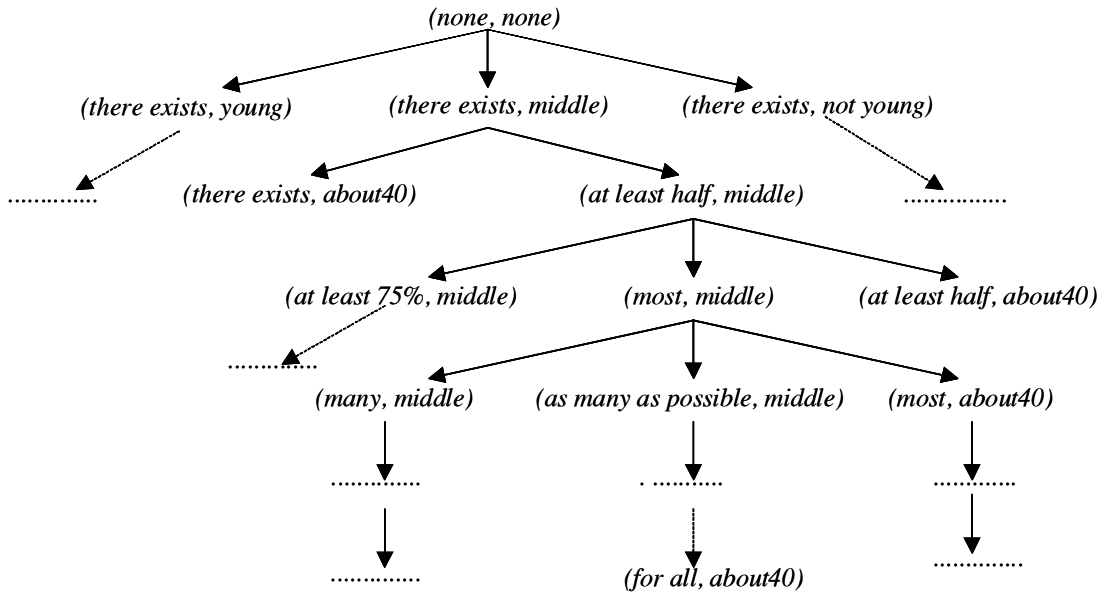
Trong đó:

$CS, SS, NextCS$ là các mảng chứa các cặp (Q, F).

Sat(H, α, β) là thủ tục kiểm tra bộ $H = (Q, F)$ có thỏa luật mẫu với ngưỡng α, β hay không.

Child(H) là tập các nút con của nút H trong cây phân cấp của luật mẫu.

Với các lượng từ unimodal ta có kết quả sau:



Hình 2. Cây phân cấp luật "*Q Age is F*"

Định lý 4.1. *Nếu Q là một lượng từ unimodal được phân tích thành hai lượng từ đơn điệu $Q = Q_a \cap Q_d$, thì*

$\Pi_{Q\ A\ is\ F}(r) \geq \alpha$ *khi và chỉ khi* $\Pi_{Q_a\ A\ is\ F}(r) \geq \alpha$ *và* $\Pi_{Q_d\ A\ is\ F}(r) \geq \alpha$

$N_{Q\ A\ is\ F}(r) \geq \beta$ *khi và chỉ khi* $N_{Q_a\ A\ is\ F}(r) \geq \beta$ *và* $N_{Q_d\ A\ is\ F}(r) \geq \beta$

Chứng minh. Dễ dàng. ∎

Cho tập các lượng từ , ta có thể phân hoạch thành 3 tập $_a$ là tập các lượng từ đơn điệu tăng, $_d$ là tập các lượng từ giảm và $_u$ là tập các lượng từ unimodal. Giả thiết là tập các lượng từ đóng đối với các lượng từ unimodal, nghĩa là với mọi lượng từ $Q \in$ $_u$ đều tồn tại $Q_a \in$ $_a$ và $Q_d \in$ $_d$ sao cho $Q = Q_a \cap Q_d$.

Từ định lý trên ta có thể hoàn chỉnh thuật toán xây dựng các luật dựa vào luật mẫu dạng 1 cho các lượng từ dạng unimodal.

Thuật toán 4.2. Xây dựng tập các luật từ luật mẫu dạng 1

Input : r là một quan hệ mờ

Luật mẫu $RL =$ "Q A *is* F"

(, $\subseteq$) là tập các lượng từ mờ với thứ tự phân cấp

(, $\subseteq$) là tập các tập mờ trên thuộc tính A với thứ tự phân cấp

Ngưỡng xác định độ tin cậy α, β

Output: Tập các cặp (Q, F) thỏa $\Pi_{Q\ A\ is\ F}(r) \geq \alpha$ và $N_{Q\ A\ is\ F}(r) \geq \beta$.

Format : DataSummary1(RL, , , α, β)

Method:

$_a$:=Rules(RL, $_a$, , α, β);

$_d$:=Rules(RL, $_d$, , α, β);

$_u$:=$\emptyset$;

For each $Q = Q_a \cap Q_d$ in $_u$ **do**

For each F in **do**

If $(Q_a, F) \in$ $_a$ and $(Q_d, F) \in$ $_d$ **then**

$_u$:= $_u \cup \{(Q, F)\}$;

EndIf

EndFor

EndFor

:= $_a \cup$ $_d \cup$ $_u$;

Return ;

Tương tự trên, các luật mẫu dạng 2 và 3 hoàn toàn có thể xây dựng thuật toán xây dựng tập luật dựa vào thứ tự phân cấp các tập mờ. Ngoài ra các kết quả trên vẫn còn đúng khi chúng ta mở rộng các tân từ nguyên tố bởi sự kết hợp các tân từ nguyên tố với phép toán and.

5. KẾT LUẬN

Với một số kết quả mở rộng tính toán về các luật tổng kết dữ liệu có yếu tố mờ đã cho phép chúng ta đánh giá được độ tin cậy của một số luật thường gặp trong thực tế và điều này là cần thiết cho các nghiên cứu tiếp theo. Việc phát hiện các luật dữ liệu với sự hỗ trợ của các chuyên gia đã được thực hiện bước đầu qua Thuật toán 4.1. và 4.2. và có thể mở rộng cho nhiều dạng luật khác. Các nghiên cứu tiếp theo của chúng tôi sẽ hoàn chỉnh và bổ sung thêm các dạng luật khác. Những kết quả tính toán sẽ được tiếp tục nghiên cứu để cài đặt trên mô hình CSDLQH mờ mà chúng tôi đã xây dựng trên PROLOG.

Lời cảm ơn

Tác giả xin chân thành cảm ơn PGS. TS Hồ Thuần và PGS. TS Đặng Huy Ruận đã đóng góp những ý kiến quý báu trong quá trình hoàn thành bài báo này.

TÀI LIỆU THAM KHẢO

[1] Bosc P., Lietard L., Pivert O., Quantified statements and Database Fuzzy querying. *Fuzziness in Database Management Systems*, Bosc P., Kacprzyk J. eds., Physica Verlag, 1995, 275–308.

[2] Bosc P., Prade H, An introduction to the fuzzy set and possibility theory-based treatment of soft queries and uncertain or Imprecise databases. *Uncertainty Management in Information Systems : From needs to Solutions*, A. Motro, Ph. Smets, Eds., Kluwer Academic Publ., 1997, 285–324.

[3] Cubero J.C., Medina J.M., Pons O., Vila M.A., Data summarization in relational databases througt fuzzy dependencies, *Information Sciences*, **121** (1999) 233–270.

[4] Dubois D., Prade H., *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York, 1988.

[5] Dubois D., Prade H., Using fuzzy sets in database systems: Why and how?, *Proceedings of the 1996 Workshop on Flexible Query-Answering Systems* (FQAS'96), Christiansen H., Larsen H. L., Andreasen T., eds., 1996, 89-103.

[6] Dubois D., Prade H., Fuzzy cardinality and the modeling of imprecise quantification, *Fuzzy Sets and Systems*, **16** (1985) 199 – 230.

[7] Kacprzyk J., Ziólkowski A., Database Queries with Fuzzy Linguistic Quantifiers, *IEEE Transactions on Systems, Man and Cybernetics*, **16** (3) (1986) 474 – 479.

[8] Petry F., Bosc P., *Fuzzy Databases: Principles and Applications*, Kluwer, Norwell, MA, 1996.

[9] Rasmussen D., Yager R.R., SummarySQL - A Fuzzy Tool for Data Mining, *Intelligent Data Analysis*, **1**(1) (1997).

[10] Rasmussen D., Yager R.R., Finding fuzzy and gradual functional dependencies with SummarySQL, *Fuzzy Sets and Systems*, **106** (1999) 134–142.

[11] L.T.Vuong, H. Thuan, A relational databases extended by application of Fuzzy set theory and linguistic variables, *Computers and Artifical Intelligence*, **8** (2) (1989) 153– 168.

[12] H. Thuan, T. T. Thanh, On the Functional Dependencies and Multivalued Dependencies in Fuzzy relational databases, *Journal of Computer science and Cybernetics*, **17** (2) (2001) 13–19.

[13] H. Thuan, T. T. Thanh, Fuzzy Functional Dependencies With Linguistic Quantifiers, *Journal of Computer science and Cybernetics*, **18** (2) (2002) 97–108.

[14] T.T.Thành, Ngôn ngữ hỏi mềm dẻo trong các cơ sở dữ liệu quan hệ mờ, *kỷ yếu Hội nghị khoa học kỷ niệm 35 năm thành lập Học viện Kỹ thuật quân sự*, Hà nội 10/2001, 116–122.

[15] Ullman J. D., *Principles of Database systems*, Comp. Science Press, 1980.

[16] Yoshikane Takahashi, A Fuzzy query language for relational databases, *IEEE Transactions On Systems, Man, and Cybernetics*, **21** (6) (1991) 365–384.

[17] Yager R.R., Fuzzy Summaries in Database Mining, *Proceedings of the 11th Conference on Artifical Intelligence for Applications*, Los Angeles, 1995, 265–269.

[18] Zadeh L., Fuzzy sets as a Basis for Theory of Possibility, *Fuzzy Sets and Systems*, **13** (1978) 3–28.

[19] Zadeh L. A., A Computational Approach to Fuzzy Quantifiers in Natural Languages, *Computers and Mathematics with Applications*, **9** (1983) 149-184.

Nhận bài ngày 25 - 9 - 2002