Tạp chí Tin học và Điều khiển học, T.21, S.4 (2005), 352–360

XÁC ĐỊNH PHẦN TỬ NGOẠI LAI TRONG CƠ SỞ DỮ LIỆU QUAN HỆ

PHẠM HẠ THỦY

Trung tâm Tin học Kiểm toán Nhà nước

Abstract. The aim of this paper is to present the detection of the outliers in a relational database. Some concepts, definitions of the outlier on the constraints system in a relational database file; the definition and algorithm for detecting the outliers on the functional dependency; some examples relating in the problems on detecting the fraud and the mistake in audit activity, are also introduced.

Tóm tắt. Bài báo trình bày việc phát hiện phần tử ngoại lai trong cơ sở dữ liệu dạng quan hệ. Một số khái niệm, định nghĩa phần tử ngoại lai theo hệ ràng buộc trong file cơ sở dữ liệu quan hệ; định nghĩa và thuật toán xác định phần tử ngoại lai theo phụ thuộc hàm; một số ví dụ ứng dụng liên quan đến việc phát hiện sai sót và gian lận trong hoạt động kiểm toán cũng được giới thiệu trong nội dung bài viết.

1. GIỚI THIỆU

Công nghệ khám phá tri thức trong cơ sở dữ liệu (CSDL) đang là chủ đề nóng trong công nghệ thông tin. Các hướng nghiên cứu chính theo hướng này tập trung vào nhận dạng và phân lớp mẫu trong cơ sở dữ liệu lớn bằng máy. Xác định phần tử ngoại lai (outlier) trong tập hợp dữ liệu là một hướng mới được quan tâm nghiên cứu và tỏ ra có nhiều ứng dụng thiết thực (xem [5, 7]). Phần tử ngoại lai trong cơ sở dữ liệu gồm hai loại: loại thứ nhất là các dữ liệu được thu thập hoặc tạo sinh theo một quy luật khác với các dữ liệu khác và được xem là dữ liệu sai hay dữ liệu không hợp lệ, loại thứ hai là dữ liêu hợp lệ nhưng có những đặc điểm khác biệt so với đa số dữ liệu. Cả hai loại đều có đặc tính chung là có dấu hiệu khác biệt so với đa số các dữ liệu khác. Vấn đề đặt ra là phát triển các phần mềm để phát hiện tự động các phần tử có dấu hiệu khác biệt trong CSDL cho phép các chuyên gia xác định xem cần loại bỏ nó ra khỏi CSDL hay cần xử lý đặc biệt đối với các phần tử ngoại lai được phát hiện này. Đến nay, ngoài các phương pháp xác định dữ liệu ngoại lai bằng phương pháp thống kê, các tác giả khác đều xác định phần tử ngoại lai theo phương pháp so sánh khoảng cách hay mức tương đồng giữa các dữ liệu.

Trong thực tiễn, nhiều dữ liệu được xem là hợp lệ nếu nó thỏa mãn các luật nào đó, nếu một trong các luật này bị vi phạm thì xem là phần tử ngoại lai. Trong bài báo này, chúng tôi sẽ xác định phần tử ngoại lai trong các cơ sở dữ liệu quan hệ dựa theo những ràng buộc, luật mà các phần tử của file dữ liệu quan hệ phải tuân theo (chẳng hạn thỏa mãn phụ thuộc hàm). Khái niệm và thuật toán đề xuất được minh họa bằng một số ví dụ minh họa trong lĩnh vực kiểm toán.

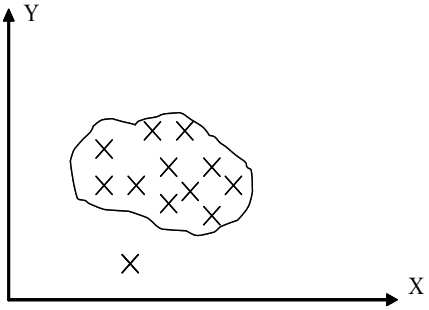
Ngoài phần kết luận, phần còn lại của bài này được trình bày như sau. Trong Mục 2, sau

khi giới thiệu tổng quan các khái niệm phần tử ngoại lai và các phương pháp tiếp cận của các tác giả khác, chúng tôi trình bày định nghĩa phần tử ngoại lai trong cơ sở dữ liệu quan hệ theo phụ thuộc hàm và theo hệ ràng buộc dạng phụ thuộc hàm. Mục 3 giới thiệu một thuật toán xác định phần tử ngoại lai đã được đề xuất. Mục 4 dành để giới thiệu một số ví dụ ứng dụng kết quả nghiên cứu ở trên để phát hiện sai sót, gian lận trong lĩnh vực kiểm toán.

2. KHÁI NIỆM VÀ ĐỊNH NGHĨA PHẦN TỬ NGOẠI LAI

2.1. Phần tử ngoại lai là gì?

Một cách hình thức người ta có thể định nghĩa phần tử ngoại lai của một tập dữ liệu là các phần tử mà theo một cách nhìn nào đó có các đặc tính không giống với tập hợp đa số còn lại của tập dữ liệu. Chẳng hạn, Hình 1 cho thấy một phần tử ngoại lai theo vị trí hình học.



Hình 1. Phần tử ngoại lai trong tập điểm có tọa độ (x, y) trên mặt phẳng có giá trị tung độ y nhỏ hơn hẳn các phần tử khác của tập hợp

Các khái niệm về ngoại lai đầu tiên có nguồn gốc từ lĩnh vực thống kê. Barnett và Lewis định nghĩa: một phần tử ngoại lai là một quan trắc hoặc một tập con các quan trắc mà sự xuất hiện của chúng trái ngược với những quan trắc còn lại (xem [4]). Phần tử ngoại lai cũng có thể được hiểu như một quan trắc mà giá trị của nó khác biệt quá nhiều so với những quan trắc khác gây cho người ta nghi ngờ rằng nó đã được thực hiện bằng một kỹ thuật khác (xem [3]). Nói một cách khác, những quan trắc không tuân theo cùng mô hình thống kê như các quan trắc còn lại được coi là các phần tử ngoại lai.

Có nhiều cách định nghĩa và hiểu khác nhau về phần tử ngoại lai. Tuy nhiên chúng có điểm chung là: một phần tử ngoại lai là những quan trắc mà có sự khác biệt đáng kể đối với những quan trắc còn lại.

Có nhiều công trình nghiên cứu về phát hiện phần tử ngoại lai. Các phương pháp chính để xác định phần tử ngoại lai bao gồm:

- *Xác định phần tử ngoại lai theo khoảng cách (Distance-Based):*

Theo hướng tiếp cận này cần phải xác định một hàm đo khoảng cách (metric) giữa các phần tử trong tập dữ liệu. Các phần tử ngoại lai là những phần tử nằm khá xa với tập các phần tử còn lại. Điển hình cho hướng tiếp cận này là E. Knorr [5].

- *Xác định theo thống kê (Statistical-Based):*

Hướng nghiên cứu này dựa trên việc xác định các mô hình phân phối thống kê mà các phần tử phải tuân theo (phân phối chuẩn, phân phối χ^2...). Phần tử ngoại lai là những phần

tử không tuân theo các luật này. Điển hình cho hướng tiếp cận này là các tác giả Barnett, Lewis (xem [4]).

- *Xác định theo độ khác biệt (Deviation-Based):*

Hướng nghiên cứu này dựa trên việc xác định những đặc trưng cơ bản của các phần tử trong một tập các phần tử. Các phần tử có những đặc trưng khác biệt quá lớn so với các phần tử còn lại thì là các phần tử ngoại lai. Điển hình cho hướng tiếp cận này là các tác giả Arning, Agrawal, Raghavan ([8]).

Các phương pháp nghiên cứu trên hiệu quả khi áp dụng trong lĩnh vực Data mining (nghiên cứu phát hiện các tri thức, các luật trong một tập các phần tử dữ liệu). Tuy nhiên chúng khó áp dụng, hoặc không hiệu quả trong các trường hợp đối với các dữ liệu của cơ sở dữ liệu dạng quan hệ trong đó có nhiều thuộc tính vừa là số và vừa là định danh, hoặc trong trường hợp khi chúng ta quan tâm nhiều đến sự vi phạm của các phần tử dữ liệu đối với một *tập các ràng buộc, quy tắc (luật) được cho trước.* Ở đây chúng tôi đề xuất việc phát hiện các phần tử ngoại lai trong CSDL quan hệ dựa theo các luật (Rule - Base). Hướng tiếp cận này giúp khắc phục được những hạn chế của các hướng nghiên cứu trước đồng thời có thể mang lại hiệu quả hơn trong việc phát hiện những phần tử ngoại lai trong CSDL quan hệ.

2.2. Phần tử ngoại lai trong cơ sở dữ liệu quan hệ

Định nghĩa 1. Với một file dữ liệu quan hệ r có các phần tử buộc phải tuân theo những quy tắc (ràng buộc) nào đó. Phần tử ngoại lai là những phần tử của file dữ liệu này không tuân theo các quy tắc đó.

Các (quy tắc) ràng buộc được đề cập bao gồm những ràng buộc về cấu trúc của CSDL (phụ thuộc hàm, các dạng chuẩn phải tuân theo - khái niệm về phụ thuộc hàm, các dạng chuẩn có thể xem trong [1,2,9]) và các ràng buộc theo ngữ nghĩa phụ thuộc vào yêu cầu, ý nghĩa của ứng dụng mà trong đó CSDL được sử dụng ([10]). Dưới đây, chúng tôi trình bày định nghĩa phần tử ngoại lai theo phụ thuộc hàm.

Định nghĩa phần tử ngoại lai theo phụ thuộc hàm

Giả sử cho một sơ đồ quan hệ (R, F), với lược đồ $R(A_1, A_2, ..., A_n)$ và tập các phụ thuộc hàm F đúng trên R. Gọi F^+ là tập các phụ thuộc hàm dẫn xuất từ F theo hệ tiên đề Amstrong. Giả sử cho r là một quan hệ trên (R, F).

Định nghĩa 2. Ta sẽ gọi một cặp $t_1, t_2 \in r$ không thỏa mãn điều kiện phụ thuộc hàm của F là cặp phần tử ngoại lai của quan hệ r.

Ta biểu diễn một cách hình thức như sau: Giả sử $X \to Y$ là một phụ thuộc hàm thuộc F^+. Khi đó cặp $t_1, t_2 \in r$ là cặp phần tử ngoại lai đối với phụ thuộc hàm $X \to Y$ nếu $t_1(X) = t_2(X)$ nhưng $t_1(Y) \neq t_2(Y)$.

Thuật toán xác định các cặp phần tử ngoại lai theo phụ thuộc hàm theo định nghĩa trên sẽ được chúng tôi trình bày ở một bài viết khác. Phần dưới chúng tôi xét tới một trường hợp những phần tử ngoại lai là những phần tử không tuân theo hệ ràng buộc dạng phụ thuộc hàm mà có thể coi là một trường hợp riêng của các phần tử ngoại lai theo phụ thuộc hàm và có ứng dụng trong thực tế của hoạt động kiểm toán.

Định nghĩa phần tử ngoại lai theo hệ ràng buộc dạng phụ thuộc hàm

Cho một lược đồ quan hệ $R(A_1, A_2, A_3, ..., A_n)$ và một quan hệ r trên R. Giả sử miền giá trị của A_i là D_i $(i = 1, ..., n)$. Giả thiết r có dạng chuẩn 1 trở nên. r được quy định:

a) Mọi bộ thuộc r phải thỏa tập các quy tắc $F(f_1, f_2, ..., f_m)$ có dạng

$$F : \{f_j : mj \Rightarrow u_j\},\ j = 1, ..., m,\ m \geq 1. \tag{1}$$

b) Mỗi bộ thuộc r đều phải thỏa mãn một trong các quy tắc f_j, trong đó, các mj, u_j, f_j là các mệnh đề lôgic.

$$mj = (A_{j1} = a_{j1}) \wedge (A_{j2} = a_{j2}) \wedge ... \wedge (A_{jk} = a_{jk}),$$

$$u_j = (A^*_{j1} = b_{j1}) \wedge (A^*_{j2} = b_{j2}) \wedge ... \wedge (A^*_{jk} = b_{jk}),$$

với $n \geq k \geq 1; n \geq d \geq 1; A_{j1}, ..., A_{jk}, A^*_{j1}, A^*_{j2}, ..., A^*_{jd} \in R$; $a_{j1}, a_{j2}, ..., a_{jk}, b_{j1}, b_{j2}, b_{jd}$ thuộc miền giá trị tương ứng của $A_{j1}, ..., A_{jk}, A^*_{j1}, A^*_{j2}, ..., A^*_{jd}$.

Cũng cần lưu ý rằng chúng ta có thể loại khỏi hệ ràng buộc (1) những quy tắc có thể suy ra từ các quy tắc khác theo các luật suy diễn của lô gic mệnh đề để biến đổi hệ quy tắc (1) thành một hệ quy tắc tối thiểu (trong đó không có các mệnh đề có thể suy diễn từ các mệnh đề khác).

Gọi $F_r = \{(A, B) : A, B \subseteq R, A \rightarrow B\}$ là họ đầy đủ các phụ thuộc hàm của r.

Ký hiệu F_r^+ là bao đóng của F_r (họ tất cả các phụ thuộc hàm có thể suy dẫn từ F_r theo hệ tiên đề Amstrong). Ký hiệu, $Bj = (A_{j1}, A_{j2}, ..., A_{jk}), Qj = (A^*_{j1}, A^*_{j2}, ..., A^*_{jd})$, chúng ta xét tập các phụ thuộc hàm:

$$G\{g_j : Bj \rightarrow Qj)\} \text{ với } j = 1, ..., m.$$

Ta có $G \subseteq F_r$.

Việc biến đổi hệ quy tắc (1) trở thành hệ tối tiểu có ý nghĩa quan trọng trong việc làm giảm số lượng tính toán trong các thuật toán kiểm tra về sau, sẽ được trình bày ở những công trình sau.

Định nghĩa 3. Một phần tử thuộc quan hệ r được gọi là phần tử ngoại lai theo ràng buộc dạng phụ thuộc hàm nếu không thỏa mãn a) hoặc b) được nêu ở trên (hệ 1).

Tách quan hệ r thành các tập con Sj sao cho mỗi tập con Sj chứa các bản ghi thỏa mãn vế trái của quy tắc f_j, nghĩa là:

$$Sj = \{t \in r, \text{ sao cho } (A_{j1} = a_{j1}, A_{j2} = a_{j2}, ..., A_{jk} = a_{jk})\},$$

$$r = \cup Sj\ (j = 1, ..., m).$$

Ta cũng dễ chứng minh bổ đề sau.

Bổ đề 1. *Đối với mỗi tập con Sj, nếu có một phần tử thỏa mãn quy tắc f_j thì khi đó phụ thuộc hàm g_j đúng trên Sj khi và chỉ khi Sj thỏa quy tắc f_j.*

Từ đây chúng ta thấy rằng trường hợp phần tử ngoại lai theo ràng buộc dạng phụ thuộc hàm sẽ là trường hợp riêng của phần tử ngoại lai theo phụ thuộc hàm.

3. XÂY DỰNG THUẬT TOÁN XÁC ĐỊNH PHẦN TỬ NGOẠI LAI THEO RÀNG BUỘC DẠNG PHỤ THUỘC HÀM

Để xác định phần tử ngoại lai trong một file dữ liệu quan hệ r thỏa mãn hệ quy tắc (1), việc đầu tiên là ta phải tách file dữ liệu này thành các phần mà trong mỗi phần đó các bản ghi thỏa mãn vế trái của quy tắc f_j trong hệ quy tắc (1).

Thực hiện việc tách nói trên chính là việc thực hiện các phép chọn trên r thỏa mãn điều kiện: $mj = (A_{j1} = a_{j1}) \wedge (A_{j2} = a_{j2}) \wedge ... \wedge (A_{jk} = a_{jk})$

$$Sj = \sigma_{mj}(R),\ j = 1, ..., m.$$

Tiếp theo, trong mỗi phần CSDL được tách ra đó ta lần lượt thực hiện (trường hợp không tồn tại phần tử nào thỏa mãn quy tắc f_j thì cả tập con các phần tử đó là phần tử ngoại lai):

+ Trong mỗi tập con Sj, kiểm tra thỏa mãn mệnh đề lôgic u_j (vế phải của (1))

$$u_j = (A^*_{j1} = b_{j1}) \wedge (A^*_{j2} = b_{j2}) \wedge ... \wedge (A^*_{jk} = b_{jk}).$$

+ Các phần tử không thỏa mãn u_j là các phần tử ngoại lai.

Dưới đây là thuật toán xác định phần tử ngoại lai theo ràng buộc dạng phụ thuộc hàm của quan hệ r.

Thuật toán 1. Tách quan hệ r theo hệ các mệnh đề vế trái của hệ quy tắc (1).

Input: Quan hệ r; Hệ ràng buộc $\{mj\}$ - vế trái của hệ quy tắc (1).

Output: $\{Sj\}$ các tập phần tử của r thỏa mãn $\{mj\}$.

```
Begin
    For j = 1 to m do
        Sj = Φ;
        Gán mj = (A_j1 = a_j1, A_j2 = a_j2...A_jk = a_jk);
            For mỗi phần tử t ∈ r do
            If t thỏa mãn mj do
                Sj = Sj ∪ t;
            Endif;
        EndFor;
    EndFor;
End.
```

Đánh giá độ phức tạp tính toán của Thuật toán 1:

Thuật toán 1 thực tế là thuật toán chọn (Select) đối với quan hệ r theo các mệnh đề lôgic $mj : Sj = \sigma_{mj}(R)$.

Thuật toán này có độ phức tạp: $Tn = O(n)$ với n là số phần tử của r, cho một giá trị thuộc tính trong mj (xem [10]).

Nếu gọi M là số thuộc tính lớn nhất trong các mj, ta có độ phức tạp tính toán tồi nhất của Thuật toán 1 là: $Tn = O(m.n.M)$;

Thuật toán 2. Xác định phần tử ngoại lai từ các tập Sj. Kiểm tra thỏa mãn quy tắc f_j.

Input: $\{Sj\}$ - các tập con của r tách ra theo Thuật toán 1;

$\{u_j\}$ - các mệnh đề lôgic ở vế phải của hệ (1).

Output: O - tập các phần tử ngoại lai theo phụ thuộc hàm.

Begin

$O = \Phi$;

For $j = 1$ to m do

For mỗi phần tử t thuộc Sj do

If t không thỏa mãn u_j then

$O = O \cup t$;

Endif;

EndFor;

EndFor;

End.

Xác định độ phức tạp tính toán của Thuật toán 2:

Thực tế ta cũng thấy Thuật toán 2 là thuật toán chọn đối với các $\{Sj\}$ theo các mệnh đề lôgic $\{u_j\}$. Ta có:

$$O = \bigcup_{j=1,\ldots,m} \sigma_{uj}(Sj)$$

Gọi M là số lớn nhất các thuộc tính có mặt trong các mj, u_j

Gọi n là số phần tử có trong quan hệ r.

Gọi Nj là số phần tử có trong tập Sj.

Vì với mỗi phép chọn $\sigma_{uj}(Sj)$ có độ phức tạp tính toán $O(m.M.Nj)$.

Do vậy độ phức tạp tính toán của Thuật toán 2:

$$Tn = O(m.M.\Sigma Nj)) = O(m.n.M),$$

trong đó:

n - số phần tử của r;

m - số lượng quy tắc f_j trong (1);

M - giá trị lớn nhất của số các thuộc tính trong R có mặt trong mj và u_j.

Thuật toán 3. Xác định các phần tử ngoại lai theo phụ thuộc hàm.

Input: $\{Sj\}$ - các tập con của r tách ra theo Thuật toán 1;

$\{u_j\}$ - các mệnh đề lôgic ở vế phải của hệ (1);

$\{mj\}$ - các mệnh đề lôgic ở vế trái của hệ (1).

Output: O - tập các phần tử ngoại lai theo phụ thuộc hàm.

Begin

Bước 1: Thực hiện Thuật toán 1: tách r thành các tập Sj theo mj.

Bước 2: Thực hiện Thuật toán 2: kiểm tra các phần tử của Sj thỏa m ãn u_j.

End.

Tổng hợp ta sẽ có độ phức tạp tính toán tồi nhất của Thuật toán 3 là:

$$Tn = O(m.n.M) + O(m.n.M) = O(2.m.n.M),$$

trong đó, n - số phần tử của r, m - số lượng quy tắc f_j trong (1), M - giá trị lớn nhất của số các thuộc tính trong R có mặt trong mj và u_j.

4. ỨNG DỤNG TRONG KIỂM TOÁN

Trường hợp 1. Kiểm toán các chứng từ kế toán về bán hàng.

Các chứng từ kế toán là các ghi chép phản ánh các nghiệp vụ kinh tế phát sinh trong kỳ của một đơn vị (mua, bán, xuất, nhập hàng...) được lưu trữ trong các bảng dạng quan hệ. Ở dạng dữ liệu trên giấy chúng là các bảng kê, sổ chi tiết. Ở dạng dữ liệu điện tử chúng là các file dữ liệu dạng quan hệ. Thuật toán được trình bày ở trên áp dụng cho trường hợp dữ liệu điện tử. Các file dữ liệu được tạo thành do quá trình sử dụng các phần mềm kế toán hoặc bảng tính Excel của đơn vị.

Giả sử chúng ta có một file dữ liệu bao gồm các chứng từ ghi chép các nghiệp vụ kinh tế phát sinh trong kỳ. Mỗi một bản ghi là một bộ giá trị của các thuộc tính sau: **Mã Chứng từ, Mã nghiệp vụ, Mã hàng, Mã khách, Mã thuế, Ngày, Diễn giải, TKnợ,TKcó, Tỉ lệ thuế, Số tiền**.

Chẳng hạn xét một nghiệp vụ kinh tế trong kỳ: khi đơn vị bán một mặt hàng A thu một khoản tiền là 5000000đ; khách nợ tiền. Khi đó theo quy định hạch toán kế toán ta phải ghi chép như sau:

- Khách nợ tiền: phải được phản ảnh trên TKnợ = 131, TKcó = 511
- Phản ảnh xuất hàng hóa từ kho: TKnợ = 632, TKcó = 156
- Phản ảnh thuế giá trị gia tăng : TKnợ = 511, TKcó= 3331
- Với mỗi loại nghiệp vụ kinh tế sẽ có các quy tắc riêng quy định.
- Khi muốn gian lận, hoặc do sai sót, người ghi chép có thể phản ảnh sai các quy định trên (sai các quy tắc quy định) nhằm trốn thuế doanh thu hoặc làm sai lệch lượng hàng có trong kho... những chứng từ như trên phải được loại ra để xem xét.
- Vấn đề phát hiện ra được các chứng từ vi phạm nguyên tắc kế toán là một trong các hoạt động cơ bản của hoạt động kiểm toán.Với những trường hợp sai sót hoặc gian lận này chúng ta có thể ứng dụng thuật toán tìm phần tử ngoại lai theo ràng buộc dạng phụ thuộc hàm để phát hiện. Các bước tiến hành như sau:

Bước 1. Xây dựng hệ thống các ràng buộc dạng phụ thuộc hàm cho một loại hình kế toán của một doanh nghiệp hoặc đơn vị. Việc xây dựng các ràng buộc này dựa trên các quy định về hạch toán kế toán, ví dụ:

a) **Mã nghiệp vụ ⇒ TKnợ, TKcó**

Cụ thể, với mỗi nghiệp vụ kinh tế có quy định việc định khoản các tài khoản nợ, tài khoản có theo giá trị nhất định.

- Với quy định mã nghiệp vụ NV21: bán hàng, cho khách nợ tiền. Khi đó ta có ràng buộc sau:

$$\langle \text{Mã nghiệp vụ} = \text{NV21} \rangle \Rightarrow (\langle \text{TKnợ}=131 \rangle, \langle \text{TKcó}=511 \rangle)$$

- Quy định mã nghiệp vụ NV22: xuất hàng hóa trong kho, ta có ràng buộc:

$$\langle \text{Mã nghiệp vụ} = \text{NV22} \rangle \Rightarrow (\langle \text{TKnợ} = 632 \rangle, \langle \text{TKcó} = 156 \rangle)$$

Hoặc quy định liên quan đến thuế:

b) **Mã nghiệp vụ, Mã hàng ⇒ TKnợ, TKcó,Tỉ lệ thuế**

(Nghiệp vụ phát sinh và mã hàng hóa quyết định giá trị TKnợ, TKcó,Ttỉ lệ thuế)

Với quy định: - Mã nghiệp vụ NV23: trích nộp thuế VAT.

- Mã hàng: A120 - tỉ lệ thuế 10%.

Ta có ràng buộc:

$$(\langle \text{Mã nghiệp vụ=NV23}\rangle, \langle \text{Mã hàng=A120}\rangle)$$
$$\Rightarrow (\langle \text{TKnợ} = 511\rangle, \langle \text{TKcó= } 3331\rangle, \langle \text{ tỉ lệ thuế} = 0.1\rangle).$$

Việc xây dựng các ràng buộc này được căn cứ vào hệ thống tài khoản, nguyên tắc kế toán và được cụ thể hóa cho từng loại hình doanh nghiệp đơn vị.

Bước 2. Sau khi đã có hệ thống các ràng buộc, một phần mềm được xây dựng bao gồm chức năng phát hiện phần tử ngoại lai có sử dụng Thuật toán 3 được trình bày ở trên. Các chức năng phát hiện phần tử ngoại lai của phần mềm sẽ được ứng dụng vào các trường hợp cụ thể.

Trường hợp 2. Kiểm toán các chứng từ xuất nhập khẩu hàng hóa.

Trường hợp đơn vị có các nghiệp vụ xuất, nhập khẩu hàng hóa có liên quan đến tỉ lệ thuế phải nộp cho Ngân sách Nhà nước. Hiện tượng gian lận, sai sót thường xảy ra là kê khai, tính toán tỉ lệ thuế không đúng với quy định của Nhà nước. Trong trường hợp này chúng ta phải đối chiếu giữa bảng định mức thuế quy định của Nhà nước với bảng kê hàng hóa xuất, nhập khẩu của đơn vị (theo mã hàng hóa và tỉ lệ thuế) để phát hiện những chứng từ kê khai sai (phần tử ngoại lai). Chúng ta cũng giả sử rằng hai bảng trên là dạng điện tử (là hai file dữ liệu quan hệ).

Khi chúng ta kết nối hai file dữ liệu này theo khóa là mã hàng hóa, chúng ta nhận được một file dữ liệu quan hệ có ràng buộc phụ thuộc hàm giữa tỉ lệ thuế kê khai của đơn vị và tỉ lệ thuế do Nhà nước quy định (quy định phải bằng nhau). Phần tử ngoại lai trong trường hợp này là trường hợp đặc biệt của phần tử ngoại lai theo ràng buộc dạng phụ thuộc hàm. Có thể áp dụng thuật toán trên trong trường hợp này. Tuy nhiên do trường hợp này có dạng đặc biệt nên chúng tôi xây dựng thuật toán riêng cho trường hợp này và sẽ được chúng tôi đề cập ở nội dung bài viết khác.

5. KẾT LUẬN

Trên đây là một số định nghĩa và cách xác định phần tử ngoại lai trong một file CSDL quan hệ dựa trên phụ thuộc hàm. Việc phát hiện phần tử ngoại lai trong cơ sở dữ liệu quan hệ còn liên quan đến nhiều vấn đề như: hệ ràng buộc tối tiểu, xác định phần tử ngoại lai theo các dạng chuẩn, xác định phần tử ngoại lai theo phụ thuộc hàm số,... cũng như việc xác định phần tử ngoại lai trong các trường hợp khác (theo ràng buộc ngữ nghĩa) sẽ được đề cập ở nội dung của các bài viết sau.

Việc ứng dụng mô hình phát hiện phần tử ngoại lai theo luật (Rule-Base) trong cơ sở dữ liệu quan hệ có ý nghĩa to lớn trong việc giải quyết nhiều bài toán thực tế như: phát hiện sự gian lận sai sót trong lĩnh vực kiểm toán là phát hiện ra những chứng từ không hợp lệ (phần tử ngoại lai) trong một tập dữ liệu lớn các chứng từ (có nhiều trường hợp lên đến hàng vạn, hàng triệu chứng từ phải kiểm toán); hoặc ngăn chặn sự sai sót trong việc xử lý dữ liệu trong lĩnh vực thiết kế cơ sở dữ liệu phân tán,v.v... Các ứng dụng nói trên đang được chúng tôi nghiên cứu để áp dụng vào hoạt động kiểm toán của Kiểm toán Nhà nước.

TÀI LIỆU THAM KHẢO

[1] Vũ Đức Thi, *Cơ sở dữ liệu - Kiến thức và thực hành*, Nhà xuất bản Thống kê, 1997.

[2] Vũ Đức Thi, *Thuật toán trong tin học*, Nhà xuất bản Khoa học Kỹ thuật, 1999.

[3] D. Hawkins, *Indentification of Outliers*, Chapman and Hall, London, 1980.

[4] V. Barnett, T. Lewis, *Outliers in Statistical Data*, John Wiley, 3^{rd} edition, 1994.

[5] E. Knorr, R. Ng, Algorithms for mining distance-based outliers in large datasets, *Proc. of the VLDB Conference*, New York, USA, September 1998, 392–403.

[6] T. Johnson, I. Kwok, Fast computation of 2-dimensional depth contours, *Proc. KDD*, 1998, 224–228.

[7] E. M. Knorr, "Outliers and data mining: finding exceptions in data", Doctor' thesis, Dept. of Computer science, University of British Columbia, 2002.

[8] A. Arning, R. Agrawal, and P. Raghavan, A linear method for deviation detection in large databases, *Proc. KDD,* 1996, 164–169.

[9] Lê Tiến Vương, *Nhập môn cơ sở dữ liệu quan hệ*, Nhà Xuất bản Khoa học và Kỹ thuật, 1995.

[10] Tamer Ozsu M. Partrick Valduriez, *Nguyên lý các hệ cơ sở dữ liệu phân tán*, Trần Đức Quang dịch, Nhà Xuất bản Thống kê, 1999.

Nhận bài ngày 12 - 4 - 2005
Nhận lại sau sửa ngày 7 - 12 - 2005