Tạp chí Tin học và Điều khiển học, T.25, S.4 (2009), 345–354

PHÁT HIỆN LUẬT KẾT HỢP VỚI RÀNG BUỘC MỤC DỮ LIỆU ÂM

CÙ THU THUỶ[1], ĐỖ VĂN THÀNH[2]

[1]*Khoa Hệ thống Thông tin Kinh tế - Học viện Tài chính*
[2]*Bộ Kế hoạch và Đầu tư*

Abstract. The aim of this paper is to propose and to solve a problem for Mining Association Rules with Negative Item Constraints from transactional databases. Finding frequent item sets from transactional databases with Negative Item Constraints is transformed to one from transactional databases having Negative Items and it is performed via the original transactional database. An algorithm of this measure is developed based on the CHARM.

Tóm tắt. Mục đích của bài báo này là đề xuất và giải quyết bài toán phát hiện luật kết hợp với ràng buộc mục dữ liệu âm từ các cơ sở dữ liệu tác vụ. Việc tìm tập phổ biến với các ràng buộc mục dữ liệu âm từ cơ sở dữ liệu tác vụ được đưa về tìm tập phổ biến từ cơ sở dữ liệu có mục dữ liệu âm và được thực hiện thông qua cơ sở dữ liệu ban đầu. Thuật toán thể hiện giải pháp được đề xuất dựa trên việc phát triển thuật toán CHARM.

1. ĐẶT VẤN ĐỀ

Phát hiện luật kết hợp là một trong những hướng nghiên cứu quan trọng của khai phá dữ liệu (data mining), hiện đã và đang hình thành nhiều cách tiếp cận nghiên cứu và ứng dụng mới liên quan đến vấn đề này. Quá trình phát hiện luật kết hợp được thực hiện thông qua hai giai đoạn, ở đó mục đích của giai đoạn đầu là tìm các tập mục dữ liệu có độ hỗ trợ lớn hơn hoặc bằng một giá trị chung minSup nào đó cho trước và được gọi là tập phổ biến, còn của giai đoạn 2 là tìm các luật kết hợp từ các tập phổ biến tìm được ở giai đoạn 1, có độ tin cậy lớn hơn hoặc bằng một giá trị chung minConf cho trước khác. Trong quá trình đó giai đoạn 1 là phức tạp và tốn nhiều chi phí nhất [1, 12, 16].

Người ta đã chỉ ra rằng việc tìm các tập phổ biến có độ hỗ trợ cực tiểu chung như nhau trong nhiều trường hợp là không hợp lý [5, 7, 9, 15]. Để khắc phục hạn chế này hiện có 5 cách tiếp cận chủ yếu là: Tìm các tập phổ biến trong mối quan hệ có sự ràng buộc về độ hỗ trợ (1); Gắn trọng số vào mỗi mục dữ liệu để đo vai trò quan trọng của từng mục dữ liệu (2); Tìm các tập phổ biến theo độ hỗ trợ cực tiểu khác nhau tuỳ thuộc vào từng mức khái niệm của các tập mục dữ liệu (3); Tìm các tập phổ biến có độ hỗ trợ cực tiểu khác nhau (4); Tìm các tập dữ liệu với các ràng buộc (5). Ưu nhược điểm của 4 cách tiếp cận đầu đã được trình bầy trong [7], còn cách tiếp cận cuối cùng nhằm phát hiện những luật kết hợp với các ràng

buộc khác nhau tuỳ theo mục đích của người sử dụng.

Việc phát hiện luật kết hợp với ràng buộc nhằm mục đích giảm tập luật kết quả và hạn chế được rất nhiều luật dư thừa. Hướng nghiên cứu này sớm được nhiều tác giả quan tâm và hiện đã thu được nhiều kết quả nghiên cứu, ứng dụng [3, 4, 8, 10, 13, 14]. Tuy nhiên thực tế cũng cho thấy rằng giữa các mục dữ liệu còn tồn tại nhiều kiểu ràng buộc khác nữa, chẳng hạn có thể xẩy ra trường hợp có một số nhóm mục dữ liệu không bao giờ xuất hiện đồng thời trong cùng một tác vụ, nói cách khác nếu một nhóm mục dữ liệu đã xuất hiện trong một tác vụ nào đó thì có thể có nhóm mục dữ liệu khác không thể xuất hiện trong tác vụ này. Bài báo gọi kiểu ràng buộc mục dữ liệu này là ràng buộc mục dữ liệu âm.

Ví dụ thực tiễn công tác điều hành các hoạt động thương mại cho thấy trong rất nhiều trường hợp nhà nước cho phép nhập khẩu nhóm mặt hàng này, thì đồng thời phải cấm nhập khẩu nhóm mặt hàng khác; hoặc khi xây dựng các dòng thuế cho các nhóm ngành hàng, vẫn thường xẩy ra trường hợp việc cho phép tăng, giảm thuế một số mặt hàng trong nhóm phải được gắn liền với việc không cho phép tăng, giảm thuế của một số mặt hàng khác; đặc biệt trong y học thì những tình huống như vậy là khá phổ biến, chẳng hạn khi người bệnh có một số triệu chứng biểu hiện của một căn bệnh nào đó thì chắc chắn người này không thể có một số triệu chứng biểu hiện cho một số căn bệnh khác,...

Thực tiễn ấy đã nẩy sinh vấn đề cần tìm tập phổ biến và các luật kết hợp có ràng buộc mục dữ liệu âm.

Mục đích của bài báo này là nghiên cứu đề xuất vấn đề và giải pháp tìm các tập phổ biến từ cơ sở dữ liệu (CSDL) tác vụ với ràng buộc mục dữ liệu âm thông qua việc đi tìm các tập phổ biến từ CSDL tác vụ có mục dữ liệu âm bằng cách chỉ dựa vào tìm các tập phổ biến từ CSDL tác vụ ban đầu. Ở đây mục dữ liệu âm là cách biểu thị mục dữ liệu không thể xuất hiện trong một tác vụ nào đó và ta cũng qui ước gọi mục dữ liệu thông thường là mục dữ liệu dương.

Bài báo sẽ giải quyết bài toán: Tìm các luật kết hợp: $A \rightarrow B$ với:

$$support(A \cup B) \geq \text{minSup};$$
$$confidence(A \rightarrow B) \geq \text{minConf}$$

và trong điều kiện tồn tại một số ràng buộc mục dữ liệu âm.

Bài báo gồm 4 mục. Sau phần đặt vấn đề, Mục 2 sẽ chỉ ra tồn tại mối quan hệ giữa CSDL tác vụ chỉ có mục dữ liệu dương với các ràng buộc mục dữ liệu âm và CSDL tác vụ có mục dữ liệu âm, từ đó cho thấy việc tìm tập phổ biến từ CSDL tác vụ các mẫu dương trong điều kiện tồn tại các ràng buộc mục dữ liệu âm chính là bài toán tìm tập phổ biến từ CSDL tác vụ có mục dữ liệu âm, đồng thời làm rõ cơ sở lý luận của giải pháp tìm tập phổ biến từ CSDL tác vụ có mục dữ liệu âm bằng cách chỉ cần thông qua CSDL tác vụ các mục dữ liệu dương tương ứng. Mục 3 chủ yếu nhằm trình bầy thuật toán, cụ thể hoá của giải pháp nêu trên và chứng minh tính đúng đắn của nó. Cuối cùng là phần kết luận và định hướng nghiên cứu.

2. TẬP PHỔ BIẾN CÓ RÀNG BUỘC MỤC DỮ LIỆU ÂM

Giả sử $\boldsymbol{I} = \{i_1, i_2, ..., i_j, ..., i_n\}$ là tập các mục dữ liệu và được gọi là tập các mục dữ liệu dương. Ký hiệu $\overline{I} = \{-i_1, -i_2, ..., -i_j, ..., -i_n\}$, ở đây $-i_j$ là ký hiệu mục dữ liệu âm của mục dữ liệu i_j, được gọi là tập các mục dữ liệu âm của $\boldsymbol{I}$. Ký hiệu $\overline{B} \subset I$ là tập mục dữ liệu âm của tập $B \subset \boldsymbol{I}$.

Ta gọi cặp $(A, \overline{B})$, với $A \subset \boldsymbol{I}$ và $\overline{B} \subset \overline{I}$ là cặp ràng buộc mục dữ liệu âm nếu mỗi khi các mục dữ liệu trong A xuất hiện trong những tác vụ nào đó thì các mục dữ liệu trong B, ở đây $A \cap B = \phi$, là không thể xuất hiện trong các tác vụ này.

Nhận xét 1: Dễ dàng thấy rằng nói chung không tồn tại mối quan hệ tập hợp giữa các cặp ràng buộc mục dữ liệu âm, cụ thể là giả sử $(A_i, \overline{B}_i),\ i = 1, 2$ là hai cặp ràng buộc mục dữ liệu âm, từ $A_1 \subseteq A_2$, không thể rút ra được quan hệ tập hợp giữa các tập $\overline{B}_i$ tương ứng và ngược lại.

Giả sử $\boldsymbol{\Omega} \subset \boldsymbol{I} \times \boldsymbol{O}$, trong đó $\boldsymbol{I}$ là tập các mục dữ liệu và $\boldsymbol{O} = \{t_1, t_2, ..., t_m\}$ là tập các tác vụ, được gọi là CSDL tác vụ và theo quy ước $\boldsymbol{\Omega}$ được gọi là CSDL tác vụ các mục dữ liệu dương.

Ký hiệu $\mathfrak{F} = \{(A_i, \overline{B}_i),\ i = 1, 2, ..., k$ (là tập tất cả các cặp ràng buộc mục dữ liệu âm cho trước.

Giả sử X là tập con bất kỳ của $\boldsymbol{I}$, ký hiệu $Y = \{x \in \boldsymbol{I} \cup \overline{I}$ nếu $x \in \boldsymbol{I}$ thì $x \in X$ hoặc nếu $x \in \overline{I}$ thì tồn tại cặp $(A_i, \overline{B}_i) \in \mathfrak{F}$ sao cho $x \in \overline{B}_i$ và $A_i \subset X\}$.

Mệnh đề 1. *Tập các tác vụ hỗ trợ X và Y xuất hiện là như nhau.*

Chứng minh:

Giả sử tác vụ $t_i \in \boldsymbol{O}$ hỗ trợ tập X, khi đó với mọi $y \in Y$ nếu $y \in X$ thì hiển nhiên t_i chứa y, nếu không phải như vậy thì tồn tại cặp ràng buộc mục dữ liệu âm $(A_i, \overline{B}_i)$ sao cho $y \in \overline{B}_i$ và $A_i \subset X$. Do t_i hỗ trợ A_i và theo định nghĩa của cặp ràng buộc mục dữ liệu âm, t_i hỗ trợ $A_i \cup \overline{B}_i$, từ đó suy ra t_i hỗ trợ y hay nói cách khác tập t_i hỗ trợ Y.

Ngược lại, với mỗi $t_i \in \boldsymbol{O}$ hỗ trợ tập Y, với mọi $x \in X$, do $x \in Y$ nên t_i hỗ trợ x và vì vậy t_i hỗ trợ tập X. ■

Mệnh đề 2. *Bài toán tìm tập phổ biến từ CSDL tác vụ $\boldsymbol{\Omega}$ với tập các điều kiện ràng buộc mục dữ liệu âm $\mathfrak{F}$ cho trước có thể được đưa về bài toán tìm tập phổ biến từ CSDL tác vụ có mục dữ liệu âm thích hợp. Ngược lại chưa chắc đúng.*

Chứng minh:

Ký hiệu $\overline{\Omega} \subset (\boldsymbol{I} \cup \overline{I}) \times \boldsymbol{O}$ là CSDL có mục dữ liệu âm. $\overline{\Omega}$ được xây dựng từ $\boldsymbol{\Omega}$ như sau:

Duyệt theo các phần tử trong $\boldsymbol{O}$, với mỗi $t \in \boldsymbol{O}$, giả sử t hỗ trợ tập mục dữ liệu $A \subset \boldsymbol{I}$. Duyệt theo tất cả các phần tử trong $\mathfrak{F}$, nếu $\exists (A_i, \overline{B}_i) \in \mathfrak{F}$ sao cho $A_i \subset A$ thế thì ta bổ sung $\overline{B}_i$ vào A.

Theo Mệnh đề 1, giả sử X là tập phổ biến tìm được từ CSDL $\mathbf{\Omega}$ với tập ràng buộc $\mathfrak{F}$ thì Y được xác định như trong Mệnh đề 1 nêu trên sẽ là tập phổ biến đối với CSDL tác vụ có mục dữ liệu âm $\overline{\Omega}$.

Ngược lại chưa chắc đúng và sẽ được chứng minh trong Ví dụ 2 dưới đây. ■

Ví dụ 1. Giả sử $\boldsymbol{I} = \{A, B, C, D, E, F, G, H\}$ và $\overline{I} = \{-A, -B, -C, -D, -E, -F, -G, -H\}$ tương ứng là tập các mục dữ liệu dương và âm. Giả sử CSDL tác vụ $\mathbf{\Omega}$ và tập các ràng buộc mục dữ liệu âm τ tương ứng được xác định như dưới đây:

Các tác vụ	Các mục dữ liệu
t_1	BCFH
t_2	ABEF
t_3	ABCG
$t4$	ABE
$t5$	ACDEF
$t6$	BCDF

và $\mathfrak{F} = \{(AB, -D), (ABC, -E - F), (D, -G - H); (BE, -D - G)\}$.

Theo cách xây dựng $\overline{\Omega}$ như trong chứng minh Mệnh đề 2, ta nhận được CSDL tác vụ có mục dữ liệu âm như sau:

Các tác vụ	Các mục dữ liệu
t_1	BCFH
t_2	ABEF-D-G
t_3	ABCG-D-E-F
t_4	ABE-D-G
t_5	ACDEF-G-H
t_6	BCDF-G-H

Ví dụ 2. Xét CSDL tác vụ có mục dữ liệu âm $\overline{\Omega} \subset (\boldsymbol{I} \cup \overline{I}) \times O$, ở đây $\boldsymbol{I} = \{A, B, C\}$ và $\overline{I} = (-A, -B, -C\}$, như sau:

Các tác vụ	Các mục dữ liệu
T_1	AB-C
T_2	A-BC
T_3	-ABC
T_4	ABC

Bắt đầu từ tác vụ t_1, ta thấy có thể xảy ra một trong 3 cặp ràng buộc mục dữ liệu âm sau: $(A, -C); (B, -C)$ và $(AB, -C)$. Cặp đầu không thể xảy ra vì ở tác vụ t_2, A và C đồng thời xuất hiện; tương tự các cặp $(B, -C)$ và $(AB, -C)$ cũng không được chấp nhận bởi các tác vụ t_3, t_4 một cách tương ứng.

Lập luận hoàn toàn tương tự cho các tác vụ còn lại. Nói cách khác trong trường hợp này không thể xây dựng được các cặp ràng buộc mục dữ liệu âm từ CSDL tác vụ có mục dữ liệu âm. ■

Mệnh đề 3. *Giả sử X, Y được xác định như trong Mệnh đề 1. Nếu X là tập phổ biến đóng cực đại trong CSDL tác vụ $\mathbf{\Omega}$ và thoả mãn tập ràng buộc mục dữ liệu âm $\Im$ thì Y cũng là tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{\Omega}$.*

Chứng minh:

- Theo Mệnh đề 1 nếu X là tập phổ biến trong CSDL $\mathbf{\Omega}$ và thoả mãn tập ràng buộc mục dữ liệu âm $\Im$ thì Y cũng là tập phổ biến trong $\overline{\Omega}$.

- Nếu X là đóng trong CSDL $\mathbf{\Omega}$ bởi theo các phép kết nối Galois f, g, h như được xác định trong [12, 17] thì dễ dàng thấy rằng Y cũng là đóng theo các phép kết nối này trong CSDL $\overline{\Omega}$.

- Nếu tập X còn là tập cực đại trong CSDL $\mathbf{\Omega}$ thì tập Y cũng có tính chất đó. Thật vậy, giả sử $Y \cup \{y\}$ với $y \notin Y$ là tập phổ biến, khi đó với y có hai khả năng: Nếu $y \in \boldsymbol{I}$ thì $y \notin X$ và $X \cup \{y\}$ là tập phổ biến, điều này là mâu thuẫn với tính chất phổ biến cực đại của X; Nếu $y \in \overline{I}$ thì điều đó mâu thuẫn với cách cách xây dựng $\overline{\Omega}$ đó là tất cả các mục dữ liệu âm đã được xác định bởi $\Im$ và được bổ sung tối đa vào các tác vụ. ■

Nhận xét 2:

Mệnh đề 2 cho biết để tìm các tập phổ biến từ CSDL tác vụ các mục dữ liệu dương nào đó trong điều kiện có ràng buộc mục dữ liệu âm, ta có thể biểu diễn CSDL tác vụ này dưới dạng CSDL tác vụ có mục dữ liệu âm, và tập phổ biến tìm được sẽ là tập có một số mục dữ liệu âm và khi đó luật kết hợp được sinh từ các tập phổ biến này sẽ là luật có thể có mục dữ liệu âm ở một hoặc cả hai phần tiền đề và hệ quả của luật kết hợp. Người ta gọi những luật kết hợp như vậy là luật kết hợp có mục dữ liệu âm hay luật kết hợp có mẫu âm [2, 6, 11].

Nếu tập các mục dữ liệu dương không quá lớn, thì việc tìm các tập phổ biến từ CSDL tác vụ có mục dữ liệu âm có thể được thực hiện theo các thuật toán tìm tập phổ biến thông dụng như Apriori, CLOSE, CHARM, ... [1, 12, 17] bằng cách coi mỗi mục dữ liệu âm là một mục dữ liệu mới và khi đó số lượng các mục dữ liệu sau khi được bổ sung thêm có thể lớn gấp 2 lần số lượng các mục dữ liệu ban đầu.

Khi số mục dữ liệu dương là khá lớn thì giải pháp này là không khả thi vì như đã biết độ phức tạp của thuật toán tìm các tập phổ biến là hàm mũ của số các mục dữ liệu và số các tác vụ trong CSDL [16].

Thực tế, bài toán tìm các tập phổ biến từ CSDL tác vụ có mục dữ liệu âm thông qua các tập phổ biến chỉ có các mục dữ liệu dương đã được một số tác giả quan tâm nghiên cứu [2, 6, 11]. Giải pháp hiện được xem là thành công nhất về vấn đề này được giới thiệu trong [11]. Tác giả bài báo này đã đề xuất biểu diễn các tập phổ biến có mục dữ liệu âm thành 3 thành phần chỉ gồm các mục dữ liệu dương, từ đó giúp tính được độ hỗ trợ của các tập có mục dữ liệu âm và tìm tập phổ biến có mục dữ liệu âm bằng cách dựa vào cải tiến phát triển thuật toán Apriori. Tuy nhiên thuật toán tìm các tập phổ biến có mục dữ liệu âm theo cách tiếp cận này còn khá phức tạp, chưa hiệu quả và cần được nghiên cứu phát triển và hoàn

thiện tiếp.

Các Mệnh đề 2, 3 đã gợi ý rằng việc tìm các tập phổ biến đóng cực đại từ CSDL tác vụ mục dữ liệu dương $\mathbf{\Omega}$ và thoả mãn tập ràng buộc $\Im$ thực chất có thể qui được về việc tìm tập phổ biến đóng cực đại từ CSDL có mục dữ liệu âm $\overline{\Omega}$. Và việc tìm các tập phổ biến đóng cực đại có mục dữ liệu âm từ $\overline{\Omega}$ có thể được thực hiện bằng cách chỉ cần thông qua việc duyệt trên CSDL tác vụ các mẫu dương $\mathbf{\Omega}$ trên cơ sở dựa vào việc cải tiến và phát triển thuật toán CHARM. Thuật toán này hiện được xem là thuật toán tìm tập phổ biến và phát hiện luật kết hợp hiệu quả nhất [17, 18].

3. THUẬT TOÁN TÌM TẬP PHỔ BIẾN VỚI RÀNG BUỘC MỤC DỮ LIỆU ÂM

3.1. Ý tưởng thuật toán

Thuật toán được tiến hành theo hai bước khá tách biệt nhau.Trước hết sử dụng thuật toán CHARM để tìm tập phổ biến đóng cực đại với các mục dữ liệu dương từ CSDL $\mathbf{\Omega}$.

Mỗi khi tìm được tập phổ biến đóng cực đại X trong CSDL này thì khởi tạo và thực hiện bước thứ 2 bằng cách duyệt các cặp ràng buộc mục dữ liệu âm, nếu tập thứ nhất của cặp này nằm trong tập X thì bổ sung tập thứ hai của cặp ràng buộc vào một tập mà sau này sẽ trở thành tập phổ biến đóng cực đại có mục dữ liệu âm trong CSDL tác vụ $\overline{\Omega}$.

Thuật toán tìm các tập phổ biến đóng cực đại có mục dữ liệu âm được gọi là NC-CHARM.

3.2. Thuật toán NC-CHARM

NC-CHARM ALGORITHM($\mathbf{\Omega} \subset \boldsymbol{I} \times O$, minSup, $\Im$):
  Nodes = $(I_j \times g(I_j) : I_j \in \boldsymbol{I} \wedge |g(I_j)| \geq \text{minSup}\}$.
  NC-CHARM-EXTEND(Nodes, $\Im, \boldsymbol{C}$)
NC-CHARM-EXTEND(Nodes, $\Im, \boldsymbol{C}$) :
  for each $X_i \times g(X_i)$ in Nodes
    NewN = ϕ and $X = X_i$
    for each $X_j \times g(X_j)$ in Nodes, with $k(j) > k(i)$
      $X = X \cup X_j$ and $Y = g(X_i) \cap g(X_j)$
      CHARM-PROPERTY(Nodes, NewN)
      if NewN $\neq \phi$ then NC-CHARM-EXTEND(NewN)
    $Temp = X$
    while each $(A_i, \overline{B}_i) \in \Im\{$
      if $A_i \subset X$ then $X = X \cup \overline{B}_i$
      $i++\}$
    if $X = Temp$ then remove X from Nodes
    $\boldsymbol{C} = \boldsymbol{C} \cup X$ // if X is not subsumed

ở đây g là một phép kết nối Galois đã được xác định cụ thể trong [12], k là một phép sắp thứ tự nào đó cho các mục dữ liệu [17]. Hàm CHARM-PROPERTY được xây dựng trong [17].

3.3. Tính đúng đắn của thuật toán

Thuật toán NC-CHARM được xây dựng dựa trên việc phát triển thuật toán CHARM. Bước thứ nhất của thuật toán NC-CHARM sử dụng những nội dung cơ bản nhất của thuật toán CHARM để tìm tập phổ biến đóng cực đại từ CSDL tác vụ các mục dữ liệu dương. Tính đúng đắn và hiệu quả của thuật toán này đã được minh chứng trong [17].

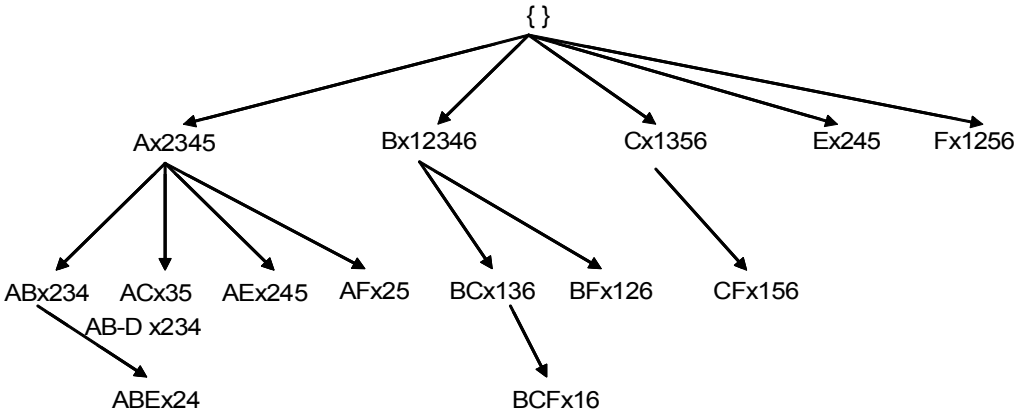
Sau khi tìm được tập phổ biến đóng cực đại X trong CSDL tác vụ các mục dữ liệu dương $\mathbf{\Omega}$, thuật toán chuyển sang bước thứ hai. Bước này, được thể hiện từ lệnh gán $Temp = X$ cho đến câu lệnh if-then cuối cùng trong thuật toán NC-CHARM, sẽ duyệt và kiểm tra các cặp ràng buộc mục dữ liệu âm xem có những mục dữ liệu âm nào cần được bổ xung tiếp vào X để tập này vẫn còn là tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{\Omega}$ hay không. Câu lệnh if-then cuối cùng sẽ bổ sung hoặc loại bỏ tập X khỏi nút của cây biểu diễn không gian tìm kiếm [17] tuỳ thuộc vào việc có hay không mục dữ liệu âm được bổ sung vào X. Tập các nút của cây, biểu diễn không gian tìm kiếm của các mục dữ liệu, $\boldsymbol{C}$ chính là tập các tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{\Omega}$.

3.4. Ví dụ minh hoạ

Xét CSDL tác vụ $\mathbf{\Omega}$ như trong Ví dụ 1 với tập các ràng buộc mục dữ liệu âm được xác định như sau: $\mathfrak{F} = \{(AB, -D), (ABC, -E - F), (D, -G - H); (BE, -D - G)\}$.

Giả sử độ hỗ trợ cực tiểu minSup $= 0,5$ và việc sắp thứ tự của các mục dữ liệu trong CSDL tác vụ các mẫu dương đã cho $\mathbf{\Omega}$ được thực hiện theo thứ tự của từ vựng. Khi đó các nút của cây không gian tìm kiếm được sắp theo thứ tự tăng dần của từ vựng.

Ban đầu khởi tạo tập Nodes = {Ax2345, Bx12346, Cx1356, Ex245, Fx1256} (dòng 1)



Hình 1. Cây tìm kiếm tập phổ biến với ràng buộc mục dữ liệu âm trên CSDL $\mathbf{\Omega}$ khi các tập mục được sắp xếp tăng dần theo thứ tự từ vựng

Vì xét theo thứ tự tăng dần của từ vựng nên thuật toán được bắt đầu ở nút Ax2345. Gán $X = A$ và kết hợp nút này với các nút lân cận phải của nó. Khi kết hợp A với B vì $g(A) \neq g(B)$ nên giữ nguyên cả hai tập và NewN = $\{AB\}$. Khi kết hợp A với C được tập AC là không phổ biến nên tỉa tập này. Khi kết hợp A với E, vì $g(E) \supset g(A)$, do vậy nhánh E sẽ bị loại bỏ, nút con AE sẽ thay thế cho E và NewN = $\{AB, AE\}$. Kết hợp A với F được tập AF không phổ biến nên cũng bị loại.

Do NewN $\neq \phi$ nên sẽ thuật toán sẽ gọi NC-CHARM cho tập này. NewN chỉ có hai phần tử. Đặt $X = AB$, sau đó kết hợp AB với AE được tập ABE không phổ biến sẽ loại bỏ.

Tiếp theo gán $Temp = \{AB\}$ và duyệt các cặp ràng buộc âm và nhận thấy có tập ràng buộc $(AB, -D)$ thoả mãn điều kiện có thành phần thứ nhất là con của tập AB vì vậy thành phần thứ hai sẽ được kết hợp vào tập $\{AB\}$ thành tập mới là $\{AB, -D\}$. Dòng lệnh tiếp theo kiểm tra thấy $X = \{AB, -D\}$ khác với $Temp$ nên bổ xung vào tập $\boldsymbol{C}$.

Tiếp tục kiểm tra điều kiện ràng buộc mục dữ liệu âm với tập $\{AE\}$, và sau đó đến $\{A\}$ nhưng cả hai tập này đều không thoả mãn điều kiện ràng buộc vì vậy đều bị loại bỏ. Kết thúc thực hiện trên nút Ax2345 ta tìm được tập $\{AB, -D\}$ là tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{\Omega}$.

Tiến hành tương tự với các nhánh Bx12346, Cx1356 và Fx1256 nhưng không tìm được tập phổ biến nào thoả mãn với tập ràng buộc mục dữ liệu âm ban đầu.

Kết thúc, ta được kết quả $C = \{AB - D\}$ là tập phổ biến đóng cực đại với ràng buộc mục dữ liệu âm.

4. KẾT LUẬN

Bài báo này đã đề xuất bài toán phát hiện luật kết hợp với ràng buộc mục dữ liệu âm từ CSDL tác vụ và đề xuất giải pháp nhằm phát hiện các luật kết hợp với những ràng buộc như vậy. Như đã biết thực chất của vấn đề phát hiện luật kết hợp từ CSDL tác vụ chủ yếu nằm ở chỗ tìm được các tập phổ biến từ CSDL này.

Trên cơ sở chỉ ra được rằng: mọi CSDL tác vụ với các ràng buộc mục dữ liệu âm đều có thể chuyển được về CSDL tác vụ có mục dữ liệu âm và bài toán tìm tập phổ biến từ CSDL tác vụ với ràng buộc mục dữ liệu âm chính là bài toán tìm tập phổ biến từ CSDL tác vụ có mục dữ liệu âm, bài báo đã đề xuất thuật toán tìm tập phổ biến đóng, cực đại từ CSDL tác vụ có mục dữ liệu âm thông qua việc tìm tập phổ biến đóng cực đại từ CSDL tác vụ chỉ có mục dữ liệu dương tương ứng.

Thuật toán đề xuất dựa trên sự cải tiến và phát triển tiếp của thuật toán CHARM, được coi là một trong những thuật toán tìm tập phổ biến (và cũng như để phát hiện luật kết hợp) từ CSDL tác vụ hiệu quả nhất hiện nay.

Như đã biết không phải CSDL tác vụ có mục dữ liệu âm nào cũng đều chuyển được về CSDL tác vụ các mục dữ liệu dương với ràng buộc mục dữ liệu âm. Nghiên cứu tiếp theo

của bài báo này sẽ là tìm các điều kiện cần và đủ để có thể thực hiện được việc chuyển đổi biểu diễn đó.

TÀI LIỆU THAM KHẢO

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, Fast discovery of association rules, *Advances in Knowledge Discovery and DataMining*, edited by U.M. fayyad, G.Platstsky-Shapiro, P.Smyth, and Uthurusamy, AAAI Press/The MIT Press, 1996 (p.306–328).

[2] M. Antonic, O. R. Zaiane, Mining positive and negative rules: an approach for confined rules, *Proc. Intl.Conf on Principles and Practice of Knowledge Discovery in Database*, Italy, 2004 (27–38).

[3] F. Bonchi, C. Lucchese, On closed constrained frequent pattern mining, *IEEE Computer Society* ICDM (2004) (35–42).

[4] C. Bucila, J. E. Gehrke, D. Kifer, and W. White, Dualminer: A dual-pruning algorithm for itemsets with constraints, *Data Mining and Knowledge Discovery* **7** (3) (2002) 241–272.

[5] C.H. Cai, "Mining Association Rules with Weighted Items", Thesis, Chinese University of Hongkong, 8/1998.

[6] C. Cornelis, P. Yan, X. Kang, G. Chen, Mining positive and negative association rules from large databases, *IEEE Computer Society*, 14244-023-6/06, 2006.

[7] Đỗ Văn Thành, Phát hiện các luật kết hợp có độ hỗ trợ cực tiểu không giống nhau, *Tạp chí Khoa học và Công nghệ* **42** (1) (2004) 79–90.

[8] L. Jia, R. Pei, and D. Pei, Tough constraint-based frequent closed itemsets mining, *Proc. Of the ACM Symposium on Applied Computing*, Melbourne, Australia, 2003.

[9] J. Han, and Y. Fu, Discovery of multiple level association rules from large databases, *Proc. of Inter. Conference on Very Large Databases*, Zurich, Swizerland, Sep. 1995 (420–431).

[10] Y. S. Koh, N. Rountree, and R.A. OKeefe, Mining interesting imperfectly sporadic association rules, *Knowledge and Information Systems* **14** (2008) 179–196.

[11] Kryszkiewicz M.; Generalized Disjunction-Free Representation of Frequent Patterns with Negation; Journal of Experimental & Theoretical Artificial Intelligence, Vol. 17, No. 1-2, pp 63-82, 2005.

[12] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Efficient mining of association rules using closed itemset latics, *Information Systems* **24** (1) (1999) 20–46.

[13] J. Pei, J. Han, and L.V.S. Lakshmanan, Mining frequent itemsets with convertible constraints, *Proc. 2001 Int. Conf. on Data Engineering (ICDE'01)*, Heidellberg, Germany, April 2001.

[14] R. Srikant, Q. Vu, and R. Agrawal, Mining association rules with item constraints, *Proceeding of the third International Conference on Knowledge Discovery and Data Mining, KDD'97*, Newport Beach, California, 2007 (67–73).

[15] W.Wang, J. Yang, P.S. Yu, Efficient mining of weighted association rules, "IBM Research Report RC 21692 (97734)", March, 2000.

[16] M. J. Zaki and M. Ogihara, Theoretical foundation of association rules, *3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Seattle, WA, USA, June 1998.

[17] M. J. Zaki, C. Hsiao, CHARM: And Efficient Algorithm for Closed Association Rule Mining, 2000, http://www.cs.rpi.edu/∼zaki.

[18] M. J. Zaki, Mining non-redundant association rules, *Data Mining and Knowledge Discovery* **9** (3) (2004) 223–248.

Nhận bài ngày 25 - 11 - 2008
Nhận lại sau sửa ngày 18 - 1 - 2010