RESEARCH ARTICLE

THUẬT TOÁN KHAI THÁC TOP-K TẬP HỮU ÍCH CAO DỰA TRÊN DI TRUYỀN VỚI ĐỘT BIẾN XẾP HẠNG

Phạm Đức Thành, Lê Thị Minh Nguyện
Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM
thanhpd@huflit.edu.vn, nguyenltm@huflit.edu.vn

TÓM TẮT— Khai thác độ hữu ích là nghiên cứu khai thác tập mục có lợi từ cơ sở dữ liệu giao dịch. Đây là phương pháp khai thác tập phổ biến dựa trên độ hữu ích để tìm tập mục phù hợp với sở thích của người dùng. Những nghiên cứu gần đây về khai thác các tập mục hữu ích cao (HUIs) từ cơ sở dữ liệu (CSDL) phải đối mặt với hai thách thức lớn, đó là không gian tìm kiếm theo cấp số nhân và ngưỡng hữu ích tối thiểu phụ thuộc vào CSDL. Không gian tìm kiếm vô cùng lớn khi số lượng các mục riêng biệt và kích thước của CSDL lớn. Các nhà phân tích phải chỉ định các ngưỡng hữu ích tối thiểu để phù hợp với những công việc khai thác của họ, mặc dù có thể họ không có kiến thức liên quan đến CSDL mà họ đang làm. Hơn nữa, thuật toán khai thác độ hữu ích chỉ hỗ trợ tập mục mang giá trị dương. Để tránh những vấn đề này, chúng tôi trình bày hai cách tiếp cận để khai thác HUI có chứa mục giá trị âm từ CSDL giao dịch: **có** hoặc **không có** chỉ định ngưỡng hữu ích tối thiểu thông qua thuật giải di truyền với đột biến được xếp hạng. Theo sự tìm hiểu của chúng tôi, đây là nghiên cứu đầu tiên trong khai thác HUI với các giá trị mang mục âm từ CSDL giao dịch sử dụng thuật giải di truyền. Kết quả thử nghiệm cho thấy các phương pháp tiếp cận được mô tả trong bài viết này đạt được hiệu suất tốt hơn về khả năng mở rộng và tính hiệu quả.

Từ khóa— Khai thác tập hữu ích cao, ngưỡng hữu ích tối thiểu, độ hữu ích, CSDL giao dịch, thuật giải di truyền.

GIỚI THIỆU

Một trong những lĩnh vực nghiên cứu chính của trí tuệ nhân tạo (AI) là khai thác dữ liệu. Sự gia tăng dữ liệu trong nhiều lĩnh vực khác nhau tạo ra một tập hợp các thách thức và cơ hội trong cách tìm kiếm và truy xuất thông tin. Bởi vì khối lượng lớn dữ liệu cần phải được phân tích, nên nhiều kỹ thuật khai thác dữ liệu được nghiên cứu áp dụng. Do đó, các nghiên cứu khai thác dữ liệu đã đang gia tăng trong những năm qua. Khai thác dữ liệu hoặc khám phá kiến thức trong CSDL (KDD), đề cập đến việc trích xuất các CSDL hợp lệ, mới lạ, có khả năng sử dụng đầy đủ, và cuối cùng là các mẫu/ kiến thức dễ hiểu trong dữ liệu [1]. Kiến thức có thể được học từ kinh nghiệm hoặc thu được từ dữ liệu. Ví dụ, thợ sửa xe thường sử dụng lý luận để tìm ra nguyên nhân thất bại và áp dụng một hành động thích hợp dựa trên kiến thức của họ về khoa học cơ khí. Các nhà phân tích kinh doanh trong một ngân hàng có thể đánh giá rủi ro thẻ tín dụng và quyết định cung cấp thẻ tín dụng cho khách hàng bằng cách phân tích dữ liệu giao dịch. Do đó, khám phá kiến thức được sử dụng để giải quyết những vấn đề phức tạp. Các phương pháp khai thác dữ liệu có thể tạo ra các loại kiến thức chẳng hạn như luật kết hợp, luật phân lớp, gom cụm và những kiến thức khác.

Vấn đề trích xuất luật kết hợp đã nhận được sự quan tâm đáng kể về nghiên cứu và nhiều thuật toán về các luật kết hợp đã được phát triển bởi Agrawal và đồng sự [2], [3], [4]. Luật kết hợp khai thác từ CSDL giao dịch là một quy trình gồm hai bước: (1) tìm kiếm tất cả các tập mục có mặt trong ít nhất s% giao dịch (tập phổ biến) và (2) phát sinh luật từ mỗi tập mục lớn [4]. Những thuật toán khai thác luật kết hợp (ARM) chỉ xem xét sự hiện diện hoặc không hiện diện các mục trong một giao dịch hoàn chỉnh; chúng không phản ánh các yếu tố ngữ nghĩa như chi phí, lợi nhuận, v.v. Các thuật toán khai thác mẫu hữu ích cao [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] giải quyết các vấn đề trong ARM bằng cách xem xét đến giá trị phi nhị phân trong các giao dịch và các giá trị lợi khác nhau của mọi mục.

Giá trị hữu ích của một mục do người dùng xác định và không có sẵn trong CSDL giao dịch. Nó phản ánh sở thích của người dùng và có thể được đại diện bởi hàm hữu ích ngoại hoặc bảng hữu ích. Bảng hữu ích xác định các hữu ích của tất cả các mục trong một CSDL nhất định. Hơn nữa, chúng ta cũng cần những hữu ích nội như số lượng mục trong giao dịch. Hàm hữu ích được biểu diễn để tính độ hữu ích của một tập mục có cả hữu ích nội và hữu ích ngoại. Xét $u(.)$là hàm hữu ích. Một tập mục I là một HUI nếu nó thỏa mãn ngưỡng $minUtil$ hoặc $u(I) > minUtil$; $minUtil$ là ngưỡng do người dùng xác định. Giá trị hữu ích của một tập mục có thể là được đo bằng chi phí, lợi nhuận hoặc các độ đo khác liên quan đến sở thích của người dùng.

Yếu tố quan trọng nhất quyết định sự thành công của sản phẩm trong tiếp thị và bán lẻ là giá của một sản phẩm. Trên thực tế, chất lượng và hình dáng bao bì đóng một vai trò quan trọng trong quyết định của khách hàng, nhưng người mua đang bị thao túng bởi giá cả, thậm chí họ thường không biết về nó. Nhiều siêu thị có thể muốn quảng cáo một số sản phẩm nhất định để thu hút khách hàng và tăng doanh số bán hàng. Chiến lược định giá lỗ để thu hút khách hàng là một loại phương pháp xúc tiến bán hàng trong đó doanh nghiệp cung cấp một mặt hàng hoặc một sản phẩm với giá thấp (hoặc miễn phí) không phải là có lợi nhuận, vì lợi ích của việc cung cấp một sản phẩm khác hoặc một mặt hàng ở mức cao hơn lợi nhuận hoặc để thu hút khách hàng mới. Trong trường hợp này,

các siêu thị có thể cho một mặt hàng làm quà tặng (tức là miễn phí) bất cứ khi nào khách hàng mua một mặt hàng cụ thể. Một mục được tặng miễn phí được coi là có giá trị âm. Loại này thực tế chủ yếu được thực hiện bởi các siêu thị để quảng bá sản phẩm, cũng như thu được lợi nhuận cao với những vật phẩm miễn phí này. Ví dụ, nếu một khách hàng mua ba món đồ I_4, người ấy sẽ nhận được một món đồ I_3 miễn phí từ siêu thị (tham khảo bảng 1). Trong trường hợp này, siêu thị thu được sáu đô la lợi nhuận từ mỗi mặt hàng I_4 đã bán và giảm hai đô la cho mỗi mặt hàng I_3 (tham khảo Bảng 2). Mặc dù siêu thị lỗ hai đô la nếu tặng miễn phí mục I_3, họ có thể kiếm được mười tám đô la khi bán ba mặt hàng I_4. Cuối cùng, họ có lợi nhuận ròng là 16 đô la từ loại quảng cáo sản phẩm này.

Các thuật toán khai thác độ hữu ích trong phần tài liệu tham khảo không trích xuất được HUI từ CSDL có mục giá trị âm. Để giải quyết vấn đề này, Chu và đồng sự [7] đã phát triển một chương trình có tên là tập mục độ hữu cao với mục giá trị âm (HUINIV-Mine) để khai thác HUI với mục giá trị âm từ CSDL lớn. Đề án HUINIV-Mine tạo ra sử dụng thuật toán hai giai đoạn [16] để khai thác HUI từ CSDL. Thuật toán TP dựa trên một ứng cử viên tạo phương pháp tiếp cận và kiểm tra, cần một lượng thời gian đáng kể để khai thác HUI.

Những thách thức lớn mà nhà phân tích dữ liệu phải đối mặt là:

- Không gian tìm kiếm cho khai thác HUI là cấp số nhân. Các yếu tố chính quyết định không gian tìm kiếm là kích thước của giao dịch và số lượng các mục trong CSDL giao dịch.
- Các nhà phân tích dữ liệu cần chỉ định ngưỡng tiện ích tối thiểu để khai thác HUI.

Có nhiều thuật toán và công nghệ để khám phá HUI đã được đề xuất bởi các nhà nghiên cứu. Những kỹ thuật này chủ yếu tập trung vào nâng cao khả năng mở rộng và hiệu quả. Các thuật toán khai thác độ hữu ích được đề xuất trong tài liệu chủ yếu dựa trên giả định rằng người dùng có thể chỉ định ngưỡng tiện ích tối thiểu phù hợp với CSDL của họ. Nhưng thiết lập ngưỡng hữu ích tối thiểu hoàn toàn không phải là một nhiệm vụ dễ dàng.

Bảng 1. *Bảng ưu đãi*

ID mục bán / số lượng	***ID mục ưu đãi / số lượng***
(I_1,2)	(I_2,1)
(I_4,3)	(I_3,1)

Bảng 2. *Bảng lợi nhuận*

Mục	I_1	I_2	I_3	I_4	I_5	I_6
Lợi nhuận	5	-3	-2	6	10	8

Để tránh những vấn đề này, các kỹ thuật dựa trên thuật giải di truyền (GA) được thiết kế để khai thác HUI từ CSDL giao dịch một cách hiệu quả. Học thuyết "Nguồn gốc của muôn loài" của Charles Darwin xuất bản năm 1859 đã nêu chi tiết cách các sinh vật phức tạp, có khả năng giải quyết vấn đề có thể được tạo ra và được cải thiện thông qua một quá trình tiến hóa của các thử nghiệm ngẫu nhiên, lai ghép và chọn lọc [17]. GA được sử dụng để xây dựng một phiên bản tiến hóa sinh học trên máy tính. GA đã được đã áp dụng thành công trong một loạt các vấn đề tối ưu hóa như kiểm soát, thiết kế, lập lịch, người máy, xử lý tín hiệu, chơi trò chơi và tối ưu hóa tổ hợp [18]. Khai thác dữ liệu cũng một trong những lĩnh vực ứng dụng quan trọng của GAs.

Những đóng góp chính của bài viết này được tóm tắt như sau:

- Cách tiếp cận tiến hóa mới được gọi là trích xuất mẫu hữu ích cao bằng cách sử dụng thuật giải di truyền với đột biến được xếp hạng sử dụng ngưỡng hữu ích tối thiểu ($HUPE_{UMU}$-GARM), sử dụng GA, được đề xuất để khai thác HUI. Trong cách tiếp cận này, nhà phân tích dữ liệu nhập vào ngưỡng hữu ích tối thiểu có giá trị ($minUtil$) cùng với CSDL giao dịch. Cách tiếp cận này được ưu tiên khi không gian tìm kiếm và việc sử dụng bộ nhớ là một vấn đề lưu tâm.
- Cách tiếp cận hiệu quả được gọi là chiết xuất mẫu hữu ích cao bằng cách sử dụng thuật giải di truyền đột biến được xếp hạng mà không sử dụng ngưỡng hữu ích tối thiểu ($HUPE_{WUMU}$-GARM) đề xuất sử dụng GA để khai thác HUI. Điều này cách tiếp cận tạo ra các HUI tối ưu mà không chỉ định một ngưỡng hữu ích tối thiểu.

Phần còn lại của bài viết này được tổ chức như sau: "Công trình liên quan". Phần tiếp theo trình bày "Những khái niệm và định nghĩa cơ bản" mô tả các khái niệm và định nghĩa cơ bản về khai thác độ hữu ích và GA. Phần kế tiếp trình bày về "Khai thác HUI với GA". Các phương pháp tiếp cận đề xuất được thảo luận trong "$HUPE_{UMU}$-GARM" và "$HUPE_{WUMU}$-GARM". Kết quả thực nghiệm là được thực hiện trong "Đánh giá thử nghiệm". Phần cuối cùng là "Kết luận".

CÔNG TRÌNH LIÊN QUAN

Trong các bản thảo gần đây dành riêng cho chủ đề khai thác độ hữu ích, chúng tôi đã xác định các yếu tố khác nhau ảnh hưởng đến hiệu suất của các thuật toán, ví dụ: cách đo độ hữu ích, cấu trúc dữ liệu và các chiến lược cắt tỉa.

Yao và đồng sự [19] phân tích lý thuyết về vấn đề khai thác độ hữu ích tối thiểu đã trình bày nền tảng cho các thuật toán khai thác độ hữu ích trong tương lai. Họ đề xuất các thuật toán giống như Apriori, được gọi là UMining và UMining_H [14] để trích xuất mức độ HUI theo từng cấp độ. Khai thác HUI thường dẫn đến việc tạo ra một số lượng lớn các mẫu. Không giống như khai thác tập mục phổ biến (FIM), mô hình khai thác độ hữu ích không thoả mãn đặc tính chống đơn điệu. Một số chiến lược cắt tỉa [20], [10], [12], [13], [14], [21], [15] đã được đề xuất để cải thiện hiệu suất của các thuật toán khai thác độ hữu ích.

Một vấn đề nghiên cứu quan trọng mở rộng từ khai thác độ hữu ích là việc phát hiện ra các mẫu hữu ích cao trong các luồng dữ liệu [11], [22], [9] vì các ứng dụng trải rộng trên nhiều lĩnh vực khác nhau. Khám phá các mẫu tuần tự hữu ích cao có thể xem như một loại hình khai thác đặc biệt, lần đầu tiên được giới thiệu bởi Yin và đồng sự [15] đã trình bày thuật toán có tên USpan để khai thác hiệu quả các mẫu tuần tự hữu ích sử dụng cây trình tự định lượng từ điển.

Các thuật toán khai thác hữu ích truyền thống không tìm thấy HUI với giá trị âm từ CSDL lớn. Vấn đề này đã được giải quyết bởi Chu và đồng sự [7], họ đã đề xuất thuật toán HUINIV-Mine. HUINIV-Mine sử dụng thuật toán TP để khai thác HUI. Khai thác quyền truy cập web trình tự (WAS) có thể trích xuất kiến thức rất hữu ích từ nhật ký web với ứng dụng rộng rãi. Ahmed và đồng sự [22] trình bày một công việc tiên phong liên quan đến khai thác WASs tiện ích cao: hai cấu trúc cây, được gọi là cây WAS dựa trên tiện ích (UWAS-tree) và cây UWAS gia tăng (IUWAS-tree) đã được đề xuất để khai thác WASs trong CSDL tĩnh và tăng trưởng.

Trong những năm gần đây, các nhà nghiên cứu tập trung vào mô hình tiếp cận khai thác duyệt web dựa trên độ hữu ích [23], [24] và xác định được các mẫu duyệt web có độ hữu ích cao. Một trong những vấn đề với phương pháp tiếp cận này là tất cả các mẫu đáp ứng ngưỡng $minUtil$ do người dùng chỉ định sẽ trở thành các mẫu độ hữu ích cao, bất kể độ dài mẫu của chúng là bao nhiêu. Vì vậy, nếu chiều dài của một mẫu tự động tăng lên, thì độ hữu ích của mẫu cũng sẽ tăng lên. Hơn nữa, các mẫu dài hơn với ít độ hữu ích hơn trong một giao dịch có thể dẫn đến giá trị cao hơn và nó sẽ được đánh giá bình đẳng với giá trị ngắn hơn các mẫu với nhiều tiện ích hơn. Thilagu và đồng sự [25] trình bày khai thác hiệu quả các mẫu duyệt web bằng cách xem xét độ hữu ích trung bình cao các mẫu hơn là các mẫu chỉ có độ hữu ích thực tế. Độ hữu ích trung bình của một mẫu có thể được định nghĩa là độ hữu ích thực tế của mẫu chia cho độ dài của nó.

Thật không may, các nghiên cứu gần đây cho thấy các thuật toán khai thác độ hữu ích có thể mắc một số điểm yếu như (1) không gian tìm kiếm lớn và (2) vấn đề với ngưỡng độ hữu ích tối thiểu phụ thuộc vào CSDL. Nhiều thuật toán khai thác HUI [5], [6], [8], [9], [11], [12], [13], [15] đã được đề xuất để giảm không gian tìm kiếm. Các thuật toán này sử dụng cách tiếp cận dựa trên cây. Nó cho thấy rằng phương pháp dựa trên cây đạt được hiệu suất tốt hơn cách tiếp cận của Apriori-like vì nó tìm thấy các HUI không có tạo các tập mục ứng viên. Liu và đồng sự [10] đề xuất thuật toán TP, khai thác hiệu quả HUI từ CSDL. Thuật toán TP sử dụng thuộc tính TWDC (bao đóng giảm) để giảm không gian tìm kiếm. Khi chúng tôi sử dụng chiến lược này thì thấy rằng nhiều ứng viên không thỏa vẫn được tạo ra. Vấn đề này có thể được giải quyết bằng cách sử dụng HUPE$_{UMU}$-GARM để khai thác HUI một cách hiệu quả bằng cách sử dụng GA.

Đặt một ngưỡng hữu ích tối thiểu phù hợp là một bài toán khó đối với các nhà phân tích dữ liệu. Nếu ngưỡng $minUtil$ được đặt quá thấp thì sẽ sinh ra một số lượng HUI lớn, điều này có thể khiến các thuật toán khai thác trở nên kém hiệu quả hoặc thậm chí hết bộ nhớ. Ngược lại, nếu ngưỡng $minUtil$ được đặt cao thì sẽ không có HUI nào được trích xuất. Chúng tôi giải quyết vấn đề này bằng cách đề xuất một phương pháp tiếp cận được gọi là HUPE$_{WUMU}$-GARM để khai thác HUI Top-K mà không chỉ định ngưỡng tối thiểu.

CÁC KHÁI NIỆM VÀ ĐỊNH NGHĨA CƠ BẢN

A. KHAI THÁC ĐỘ HỮU ÍCH

Khai thác HUI là một lĩnh vực nghiên cứu về khai thác dữ liệu mô tả dựa trên độ hữu ích, nhằm mục đích tìm kiếm các tập mục đóng góp tốt nhất vào tổng hữu ích.

Cho $I = \{I_1, I_2, \dots, I_m\}$ là một tập mục; $D = \{T_1, T_2, \dots, T_n\}$ là một CSDL giao dịch, trong đó mỗi giao dịch $T_j \in D$ là một tập con của I ; $o(I_p, T_q)$ là giá trị hữu ích giao dịch cục bộ, đại diện cho số lượng mục I_p trong giao dịch T_q . Ví dụ, $o(I_1, T_9) = 2$, trong Bảng 3. Trong bảng lợi nhuận (Bảng 2), $s(I_p)$, hữu ích ngoại, là giá trị được liên kết với mục I_p. Giá trị này phản ánh tầm quan trọng của một mục, thuộc về các giao dịch. Ví dụ, trong Bảng 2, hữu ích ngoại của mục I_1, $s(I_1)$ là 5. Độ hữu ích $u(I_p, T_q)$, thước đo định lượng của mức độ hữu ích I_p trong giao dịch T_q, được định nghĩa là $o(I_p, T_q) \times s(I_p)$. Ví dụ: $u(I_1, T_9) = 2 \times 5$ trong bảng 3. Độ hữu ích của tập mục X trong giao dịch T_q,

$u(X, Tq)$, được định nghĩa là $\sum_{I_p \in X} u(I_p, T_q)$, trong đó $X = \{I_1, I_2, ..., I_k\}$ ký hiệu là một k-itemset, $X \subseteq T_q$ và $1 \le k \le m$. Độ hữu ích của một tập mục X, $u(X)$, được định nghĩa như sau:

$$\sum_{T_q \in D \wedge X \subseteq T_q} u(X, T_q) \quad (1)$$

Nhiệm vụ chính là tìm tất cả HUI bằng cách sử dụng khai thác độ hữu ích. Một tập mục X là một tập mục hữu ích cao, nếu $u(X) \ge$ minUtil , trong đó $X \subseteq I$. Ví dụ, trong Bảng 3, $u(I_1, T_9) = 2 \times 5 = 10$, $u(\{I_1, I_4\}, T_9) = u(I_1, T_9) + u(I_4, T_9) = 2 \times 5 + 3 \times 6 = 28$ và $u(\{I_1, I_4\}) = u(\{I_1, I_4\}, T_2) + u(\{I_1, I_4\}, T_9) = 46 + 28 = 76$. Nếu minUtil = 150, thì $\{I_1, I_4\}$ không phải là HUI. Cách tiếp khai thác độ hữu ích không hỗ trợ thuộc tính bao đóng giảm [3]. Do đó, sự kết hợp của tất cả các mục được tạo ra giống nhau nên được xử lý để đảm bảo rằng không có HUI nào bị mất.

Liu và đồng sự [10] đã trình bày thuật toán TP cho khai thác HUI. Thuật toán TP có hai giai đoạn. Trong giai đoạn đầu, độ hữu ích của giao dịch (TU) cho tất cả các giao dịch được tính toán. Tiếp theo, tập hợp của tất cả tập mục đơn được xác định và tính toán độ hữu giao dịch có (TWU) cho mỗi tập mục đơn tương ứng bằng cách quét CSDL lần hai.

Bảng 3. *Bảng số lượng*

TID/ITEM	I_1	I_2	I_3	I_4	I_5	I_6
T_1	2	1	0	0	2	2
T_2	2	1	2	6	0	1
T_3	0	0	2	6	0	1
T_4	2	1	0	0	0	0
T_5	0	0	2	6	0	1
T_6	2	1	0	0	2	0
T_7	2	1	0	0	0	1
T_8	0	0	1	3	0	2
T_9	2	1	1	3	0	1
T_{10}	0	0	1	3	1	1

Sự kết hợp của các tập mục có độ hữu ích giao dịch với trọng số cao được thêm vào tập hợp các ứng viên tại mỗi cấp độ trong quá trình tìm kiếm cấp độ level-wise. Giai đoạn này duy trì một giao dịch tính chất bao đóng giảm có trọng số (TWDC) [26]. Giai đoạn đầu có thể đánh giá quá cao một số tập mục hữu ích thấp, nhưng nó không bao giờ đánh giá thấp bất kỳ tập mục nào. Trong giai đoạn thứ hai, một lần quét CSDL bổ sung là được thực hiện để lọc các tập mục được đánh giá quá cao.

Xem xét Bảng 3: có 10 giao dịch, $tu(T_1)$ là độ hữu ích giao dịch của T_1 và sẽ được tính như sau $tu(T_1) = u(I_1, T_1) + u(I_2, T_1) + u(I_5, T_1) + u(I_6, T_1) = 2 \times 5 + 1 \times (-3) + 2 \times 10 + 2 \times 8 = 43$; ($tu(T_1)$ là 46 nếu các mục chứa giá trị âm không được xem xét). Các độ hữu ích giao dịch cho tất cả các giao dịch được liệt kê trong Bảng 4. Độ hữu giao dịch có trọng số của một mục I_1, được ký hiệu là TWU(I_1), là tổng của độ hữu ích giao dịch của tất cả các giao dịch có I_1. Quan sát bảng 4 và bảng 5, TWU được tính: TWU(I_1) = TU(T_1) + TU(T_2) + TU(T_4) + TU(T_5) + TU(T_6) + TU(T_7) + TU(T_9) = 46 + 54 + 10 + 30 + 18 + 36 = 194.

Bảng 4. *Độ hữu ích giao dịch có/không có mục giá trị âm của CSDL giao dịch*

TID	TU có mục giá trị âm	TU không tính mục giá trị âm
T_1	43	46
T_2	47	54
T_3	40	44
T_4	7	10
T_5	40	44
T_6	27	30
T_7	15	18
T_8	32	34
T_9	31	36
T_{10}	34	36

Bảng 5. *CSDL giao dịch*

TID/ITEM	I_1	I_2	I_3	I_4	I_5	I_6
T_1	1	1	0	0	1	1
T_2	1	1	1	1	0	1
T_3	0	0	1	1	0	1
T_4	1	1	0	0	0	0
T_5	0	0	1	1	0	1
T_6	1	1	0	0	1	0
T_7	1	1	0	0	0	1
T_8	0	0	1	1	0	1
T_9	1	1	1	1	0	1
T_{10}	0	0	1	1	1	1

B. THUẬT GIẢI DI TRUYỀN

GA là một kỹ thuật tìm kiếm và tối ưu hóa có định hướng có thể giải quyết các vấn đề phức tạp và thường có độ phi tuyến cao. Nó được sử dụng để tìm kiếm không gian vấn đề rất lớn để tìm ra giải pháp tốt nhất dựa trên các chức năng thích nghi dưới một tập hợp nhiều ràng buộc. GA bắt đầu với một lượng lớn các giải pháp khả thi và thông qua việc áp dụng phương pháp lai ghép chéo và đột biến, phát triển một giải pháp tốt hơn bất kỳ giải pháp nào trước đây trong suốt thời gian tồn tại của phân tích di truyền. Các thuật ngữ cơ bản được sử dụng trong GA được đề cập trong bảng 6.

Bảng 6. *Các thuật ngữ được sử dụng trong thuật giải di truyền*

Thuật ngữ	**Mô tả**
Locus	Vị trí trong bộ gen được gọi là locus.
Allele	Giá trị của gen (hoặc các gen).
Genome	Đặc tính cụ thể trong nhiễm sắc thể tương ứng với bộ gen.
Chromosome	Tập các bộ gen dài được gọi là nhiễm sắc thể (cá thể)
Fitness function	Thước đo gắn liền với các chức năng mục tiêu chung cho biết sự phù hợp của một nhiễm sắc thể cụ thể.
Survival of the fittest	Những cá thể khỏe mạnh nhất được bảo tồn và sinh sản, sống sót từ những cá thể khỏe mạnh nhất.
Generation	Thế hệ là sự lặp lại của thuật giải di truyền. Thế hệ ngẫu nhiên ban đầu thường được gọi là thế hệ không.
Selection	Quá trình chọn lọc các nhiễm sắc thể tốt từ quần thể để nhân giống sau này được gọi là chọn lọc.
Crossover	Quá trình tạo ra một nhiễm sắc thể mới bằng cách lai ghép hai hoặc nhiều nhiễm sắc thể có giá trị được gọi là sự trao đổi chéo.
Mutation	Quá trình thay đổi ngẫu nhiên giá trị của gen để dẫn đến một giải pháp tối ưu được gọi là đột biến.
Search space	Không gian của tất cả các giải pháp khả thi.

KHAI THÁC CÁC TẬP HỮU ÍCH CAO VỚI THUẬT GIẢI DI TRUYỀN

Cho $I = \{I_1, I_2, \dots, I_m\}$ là một tập các mục, $D = \{T_1, T_2, \dots, T_n\}$ là một CSDL giao dịch, trong đó mỗi giao dịch $T_j \in D$ là một tập hợp con của I. Một tập mục X là một HUI nếu nó thỏa mãn ngưỡng $minUtil$ hoặc $u(X) \geq minUtil$;

minUtil là ngưỡng do người dùng xác định. Trong bài báo này, phần này được mô tả thành những lưu đồ khối trong GA.

C. MÃ HÓA

Trong phần này, mã hóa được trình bày đầu tiên. Ba toán tử di truyền và khởi tạo quần thể và hàm thích nghi dựa trên sự đo lường của Yao và đồng sự [19]. Cuối cùng, các mô-đun này được ghép lại với nhau trong thuật toán $HUPE_{UMU}$-GARM và $HUPE_{WUMU}$-GARM. Hình 1 cho biết cấu hình của các gen (mục) trong nhiễm sắc thể. Cách tiếp cận này sử dụng mã hóa nhị phân. Giá trị "1" thể hiện sự hiện diện của một mục và "0" thể hiện sự vắng mặt của một mục trong một tập mục. Nhiễm sắc thể có chiều dài là cố định và nó bằng số mục riêng biệt (n), thu được từ CSDL giao dịch.

D. KHỞI TẠO QUẦN THỂ

Với chiều dài tập mục là k, tất cả các gen (mục) trong một nhiễm sắc thể được mã hóa bằng 0. Quần thể ban đầu được tạo ra bằng cách sử dụng việc phát sinh một số ngẫu nhiên.

Hình 1. *Nhiễm sắc thể*

Nếu số ngẫu nhiên được tạo ra là *r*, thì nhiễm sắc thể được mã hóa là 1 ở vị trí thứ *r*. Điều này thể hiện mục i_r hiện diện trong một nhiễm sắc thể (tập mục). Khi tạo ngẫu nhiên một mục trong nhiễm sắc thể, nó được kiểm tra với các mục khác đã được tạo trong cùng một nhiễm sắc thể, và nếu mục có rồi, một số mới sẽ được tạo ngẫu nhiên cho đến khi là duy nhất. Điều này được lặp lại cho đến khi tạo ra k số ngẫu nhiên duy nhất, với điều kiện $k \leq n$.

Thí dụ: Giả sử rằng chiều dài nhiễm sắc thể là 7 và tập mục độ dài là 4. Các số ngẫu nhiên được tạo bởi một bộ tạo số ngẫu nhiên là {1, 3, 5, 6}.

Sự biểu diễn của một nhiễm sắc thể được thể hiện trong Hình 2.

1	**0**	**1**	**0**	**1**	**1**	**0**

Hình 2. *Biểu diễn nhiễm sắc thể cho tập mục {I1, I3, I5, I6 }*

Thuật toán khởi tạo quần thể.

Algorithm 1 **Population_Initialize**
Input Kích thước nhiễm sắc thể *n*, chiều dài tập mục *k*.
Output Quần thể mới được tạo *pop*

```
1.  i ← 0
2.  for j ← 0 to n do
3.          pop[j] ← 0
4.  while i ≤ k do
5.          rand_no ← rand(k)
6.          if pop[rand_no] ≠ 1 then
7.                  pop[rand_no] ← 1
8.                  i ← i + 1
9.          end if
10. end while
11. Trả về pop
```

E. HÀM THÍCH NGHI

Mục tiêu chính của công việc này là tạo HUI từ CSDL giao dịch CSDL. Do đó, hàm thích nghi là điều cần thiết để xác định nhiễm sắc thể (tập mục) thỏa mãn ngưỡng *minUtil*. Trong các thuật toán $HUPE_{UMU}$-GARM và $HUPE_{WUMU}$-GARM, chúng tôi sử dụng thước đo độ hữu ích của Yao Hamilton và đồng sự [19], $u(X)$ là hàm thích nghi.

$$f(X) = u(X) = \sum_{T_q \in D \wedge X \subseteq T_q} u(X, T_q) \quad (2)$$

F. TOÁN TỬ DI TRUYỀN

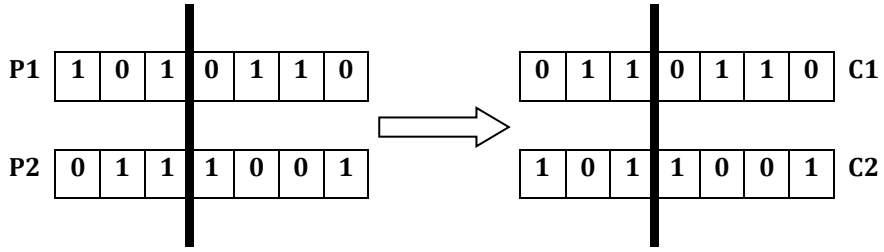
Phần phụ này thể hiện ba toán tử di truyền: chọn lọc, lai ghép và đột biến.

1. CHỌN LỌC

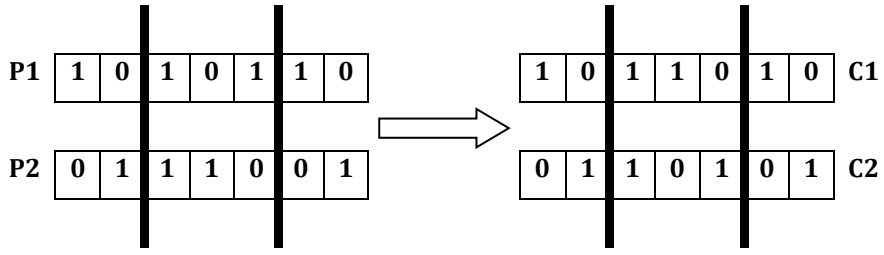
Chọn lọc là một bước của GA trong đó một cá thể được chọn từ một quần thể để tạo giống sau này (lai ghép). Nếu tất cả các gen trong nhiễm sắc thể có giá trị hữu ích âm thì nó sẽ không nhân giống sau này. Toán tử chọn lọc hoạt động như một bộ lọc của nhiễm sắc thể với những cân nhắc về độ thích nghi của chúng. Trong tài liệu tham khảo có nhiều phương pháp tiếp cận theo hướng chọn. Trong bài báo này, chúng tôi sử dụng phương pháp chọn theo bánh xe roulette của Holland [27].

2. LAI GHÉP

Lai ghép là một tính năng quan trọng của GAs. Lai ghép chọn ra hai cá thể được gọi là bố mẹ và tạo ra hai cá thể mới được gọi là con bằng cách hoán đổi các phần của cha mẹ với nhau.



Hình 3. *Lai ghép 1 điểm*



Hình 4. *Lai ghép hai điểm.*

Toán tử hoạt động đơn giản nhất là trao đổi các chuỗi con sau một điểm giao nhau được chọn ngẫu nhiên. Hình minh họa của sự giao nhau hoạt động được đưa ra trong Hình 3 và 4.

3. ĐỘT BIẾN

Toán tử đột biến được sử dụng để duy trì sự đa dạng từ một thế hệ của một quần thể đến thế hệ tiếp theo. Sự thay đổi đột biến một hoặc nhiều giá trị gen trong nhiễm sắc thể từ thế hệ trước của nó. GA đơn giản tuân theo xác suất đột biến do người dùng xác định (P_{m}) cho hoạt động đột biến. Thông thường, xác suất đột biến nên được đặt ở mức thấp. Nếu nó quá cao, tìm kiếm sẽ trở thành một tìm kiếm ngẫu nhiên ban đầu. Đột biến thích nghi tỷ lệ mang lại hiệu suất tốt hơn tỷ lệ đột biến cố định [28].

$$P_m = \left(P_m^{max} - \frac{P_m^{max} - P_m^{min}}{N_i} \times T \right) \times \frac{Rank}{R} \quad (3)$$

P_m^{max}: tỉ lệ đột biến tối đa.

P_m^{min}: tỉ lệ đột biến tối thiểu.

N_i: Số lần lặp lại

T: Số lần lặp lại

R: Tổng số bậc

Trong công trình đề xuất, nếu xác suất đột biến giảm khi số thế hệ tăng lên thì tỉ lệ đột biến của đời con thích nghi đối với giá trị thích nghi của nó. Ban đầu, để khám phá nhiều không gian tìm kiếm thì sử dụng tỉ lệ đột biến

lớn. Con cái được xếp hạng dựa trên giá trị thích nghi của nó. Tỉ lệ đột biến của đời con được gán với thứ hạng tương ứng của nó. Con cái có thứ hạng cao hơn thì tỷ lệ đột biến thấp hơn khi so với những con có thứ hạng thấp. Đời con có độ thích nghi cao nhất có thể đạt được giải pháp tối ưu trong thời gian sớm. Vì vậy, mức tỷ lệ tối thiểu tránh sự phân kỳ trong không gian tìm kiếm và P_m có thể được tính bằng Công thức (3).

G. SỰ ĐÁNH GIÁ

Bước đánh giá dự định chọn lọc các nhiễm sắc thể cho thế hệ tiếp theo. Trong bài báo này, sử dụng phương pháp chọn lọc tinh hoa [27]. Phương pháp này sao chép (các) nhiễm sắc thể có giá trị thích nghi cao hơn sang quần thể mới.

H. ĐIỀU KIỆN DỪNG

Điều kiện dừng là tiêu chí mà GA quyết định tiếp tục tìm kiếm hay dừng tìm kiếm. Điều kiện kết thúc có thể được liệt kê dưới đây:

- Số lượng thế hệ đã đạt được cố định.
- Giá trị thích nghi của giải pháp với xếp hạng cao nhất ở một con số không đổi qua các thế hệ liên tiếp.
- Kiểm tra giải pháp theo cách thủ công.
- Sự kết hợp của những điều trên.

I. THUẬT TOÁN HUPE$_{UMU}$-GARM

Khai thác HUI với ngưỡng hữu ích tối thiểu bằng GA nhằm giúp trích xuất một cách hiệu quả các mẫu hữu ích từ CSDL. Chọn GA bởi vì nó là một giải pháp đầy hứa hẹn cho tìm kiếm toàn cục và nó có khả năng khám phá HUI với tham số tương ứng là số lượng và lợi nhuận. Chúng tôi sử dụng hai độ đo TWU của Liu và đồng sự [10] để loại bỏ những mục không hứa hẹn ra khỏi CSDL giao dịch ở giai đoạn đầu và dùng độ đo của Yao và đồng sự [19], để tính toán giá trị thích nghi.

Các bước của thuật toán HUPE$_{UMU}$-GARM như sau (xem Hình 5):

Bước 1. Quét CSDL giao dịch để tính TU của mỗi giao dịch không xem xét các mục giá trị âm. Đồng thời, TWU của mỗi mục duy nhất cũng được tích lũy.

Bước 2. Nếu TWU của một mục nhỏ hơn ngưỡng hữu ích tối thiểu, tập cha của nó không hứa hẹn là HUI (thuộc tính TWDC). Vì vậy, loại bỏ các mục không hứa hẹn tương ứng từ CSDL giao dịch. CSDL giao dịch sau khi tổ chức lại ở trên được gọi là CSDL giao dịch được tái tổ chức.

Bước 3. Nhận số lượng các mục riêng biệt (NDI) từ CSDL giao dịch tái tổ chức. Đặt chiều dài nhiễm sắc thể (CL) thành NDI.

Bước 4. Tạo ra một nhiễm sắc thể có chiều dài CL (trong một quá trình tạo lập, nếu tất cả các gen trong nhiễm sắc thể có giá trị hữu ích âm thì nó sẽ không được xem xét).

Bước 5. Đánh giá từng cá thể riêng biệt bằng cách tính giá trị thích nghi (fv) và kiểm tra $fv \geq$ minUtil. Nếu có, hãy chuyển sang Bước 6, nếu không, hãy quay lại Bước 4.

Bước 6. Kiểm tra xem kích thước quần thể có bằng N không. Nếu có, hãy chuyển sang Bước 7, nếu không quay lại Bước 4.

Bước 7. Nếu Điều kiện dừng được đáp ứng, thì chỉ định cá thể tốt nhất trong quần thể như là đầu ra của quá trình tiến hóa và chấm dứt; nếu không thì tiếp tục.

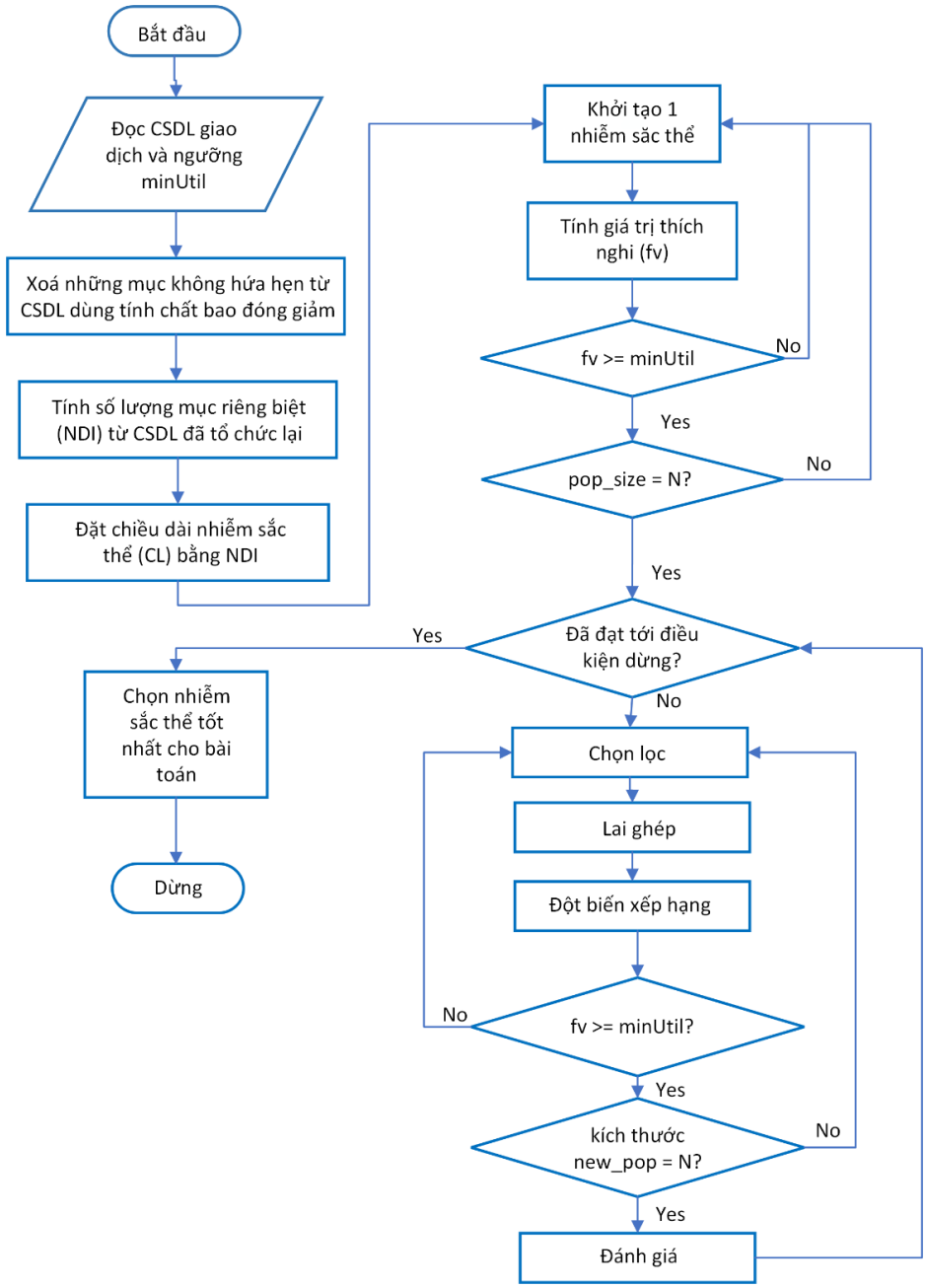
Bước 8. Chọn lọc m cá thể bằng cách sử dụng chọn lọc bánh xe roulette sẽ tạo ra thế hệ tiếp theo với những bậc cha mẹ tốt nhất.

Bước 9. Thực hiện lai ghép trên các cá thể đã chọn của quần thể.

Bước 10. Thực hiện đột biến trên các cá thể của quần thể mang đột biến xác suất P_m. Tính giá trị thích nghi (fv) cho các cá thể và kiểm tra $fv \geq$ minUtil. Nếu có, hãy quay lại bước 8, nếu không, hãy chuyển sang Bước 11.

Bước 11. Kiểm tra xem kích thước của quần thể mới có đạt đến N hay không. Nếu có, hãy chuyển sang Bước 12, nếu không, hãy quay về Bước 8.

Bước 12. Đánh giá các cá thể bằng cách sử dụng chọn lọc tinh hoa của N nhiễm sắc thể từ quần thể mới và quần thể cũ cho thế hệ tiếp theo.



Hình 5. *Lưu đồ của thuật toán HUPE$_{UMU}$-GARM.*

J. 4.9 THUẬT TOÁN HUPE$_{WUMU}$-GARM

Phương pháp tiếp cận HUPE$_{WUMU}$-GARM được đề xuất được sử dụng để khai thác tối ưu HUI mà không cần chỉ định ngưỡng *minUtil*. Trongcách tiếp cận này, dùng độ đo Yao và đồng sự [19] để tính toán giá trị thích nghi.

Các bước của thuật toán HUPE$_{UMU}$-GARM như sau (xem Hình 6):

Bước 1. Kiểm tra CSDL giao dịch để tìm số lượng các mục riêng biệt (NDI). Đặt chiều dài nhiễm sắc thể (CL) thành NDI.

Bước 2. Tạo ra một nhiễm sắc thể có chiều dài CL (trong một quá trình tạo lập, nếu tất cả các gen trong nhiễm sắc thể có giá trị hữu ích âm thì nó sẽ không được xem xét).

Bước 3. Đánh giá cá thể bằng cách tính giá trị thích nghi (*fv*).

Bước 4. Kiểm tra xem kích thước quần thể có bằng N không. Nếu có, chuyển sang Bước 5, nếu không, quay lại Bước 2.

Bước 5. Nếu tiêu chí kết thúc được đáp ứng, thì chỉ định các tập mục hữu ích Top-K từ quần thể như là đầu ra của quá trình tiến hóa và chấm dứt; nếu không thì tiếp tục.

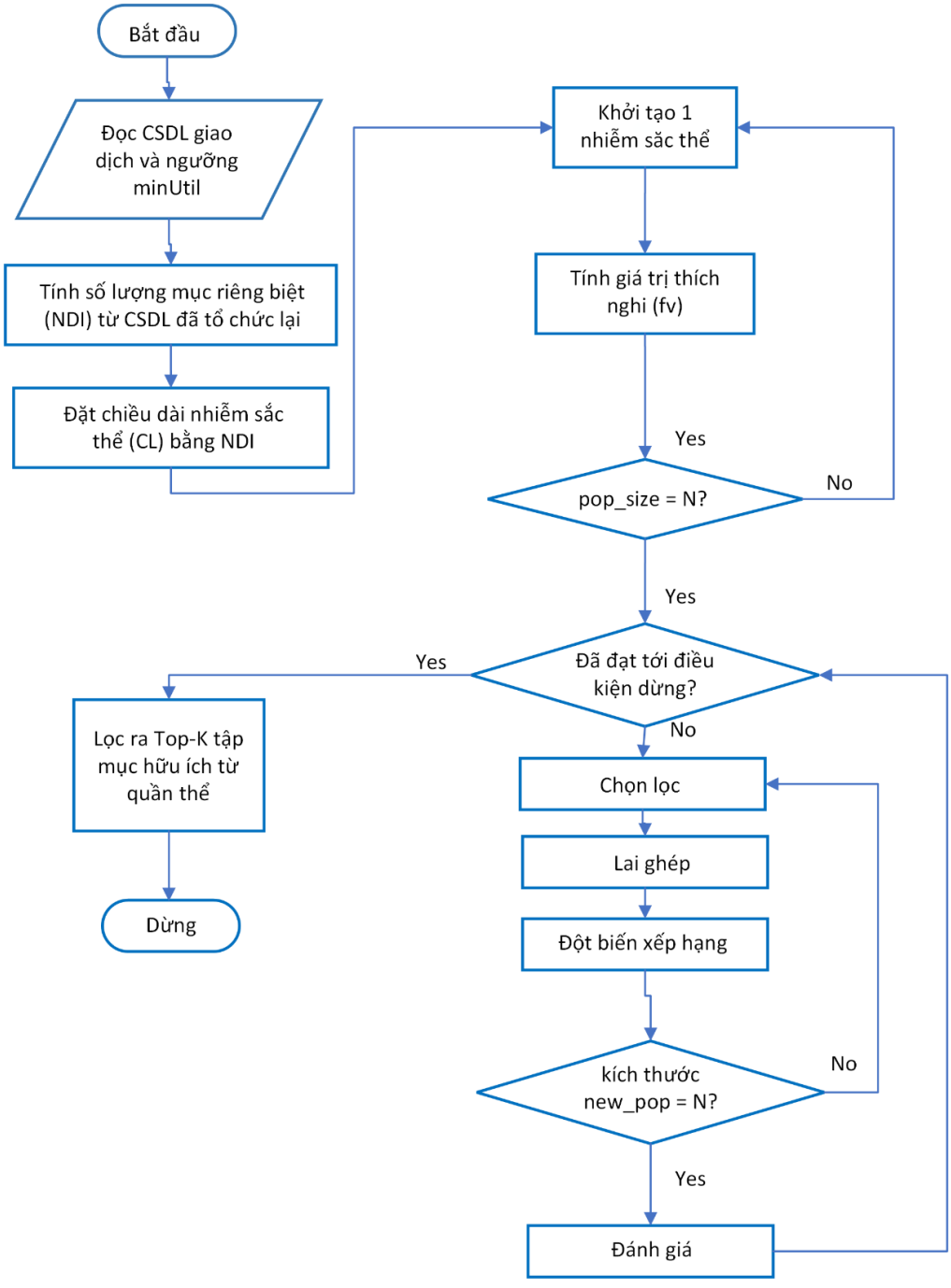
Bước 6. Chọn lọc *m* cá nhân bằng cách sử dụng chọn lọc bánh xe roulette, sẽ hợp thành thế hệ tiếp theo với những bậc cha mẹ tốt nhất.

Bước 7. Thực hiện lai ghép trên các cá thể đã chọn của quần thể.

Bước 1. 8. Thực hiện đột biến trên các cá thể của quần thể mang đột biến xác suất P_m. Tính giá trị thích nghi (*fv*) cho các cá thể.

Bước 9. Kiểm tra xem kích thước của quần thể mới có đạt N hay không. Nếu có, hãy chuyển sang Bước 10, ngược lại quay về Bước 6.

Bước 10. Đánh giá các cá thể bằng cách sử dụng sự chọn lọc tinh hoa của N nhiễm sắc thể từ quần thể mới và quần thể cũ cho thế hệ tiếp theo.



Hình 6. *Lưu đồ của phương pháp tiếp cận HUPE$_{WUMU}$-GARM.*

ĐÁNH GIÁ THỰC NGHIỆM

Thực nghiệm được thực hiện trên máy có Intel 2,90 GHz CPU Core i7-7820HQ và RAM 32GB, chạy trên Windows 10 Pro. Các thử nghiệm trong Hình 7 đến 12 được thực hiện trên bộ dữ liệu tổng hợp từ dữ liệu IBM trình tạo dữ liệu tổng hợp (IBM Quest Market-Basket Synthetic Data Generator) [29]. Dataset "T10.I4.D10K" có nghĩa là kích thước giao dịch trung bình là 10, kích thước trung bình là tập phổ biến tối đa có khả năng xảy ra là 4 và 10.000 giao dịch được tạo ra. Tập dữ liệu này chứa 100 mục riêng biệt. Số lượng cho mỗi mục trong một giao dịch cũng được giữ trong tập dữ liệu (giá trị trong khoảng từ 1 đến 10). Giá trị hữu ích cho các mục được gán ngẫu nhiên trong bảng lợi nhuận.

K. THUẬT TOÁN HUPE$_{UMU}$-GARM

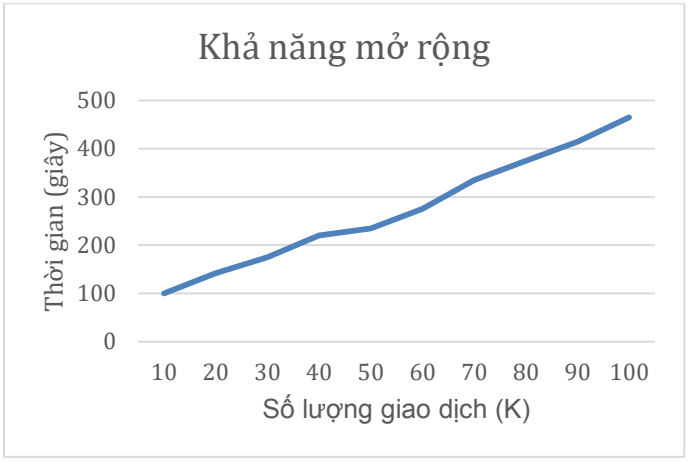
Thử nghiệm đầu tiên được thực hiện với HUPE$_{UMU}$-GARM tiếp cận ở các số lượng các thế hệ khác nhau bằng cách giữ ngưỡng $minUtil = 2\%$và 100 mục riêng biệt cũng được cố định. Kết quả kiểm tra tập dữ liệu T10.I4.D10K được minh họa qua Hình 7 và Hình 8, tương ứng.

Hình 7. *Biểu diễn số lượng mẫu tương ứng với số thế hệ với minUtil = 2%*

Hình 8. *Biểu diễn thời gian thực thi tương ứng với số thế hệ với minUtil = 2%*

Có thể nhận thấy rằng thời gian thực thi của cách tiếp cận HUPE$_{UMU}$-GARM để khai thác HUI từ CSDL được chứng minh là ngắn hơn đáng kể.

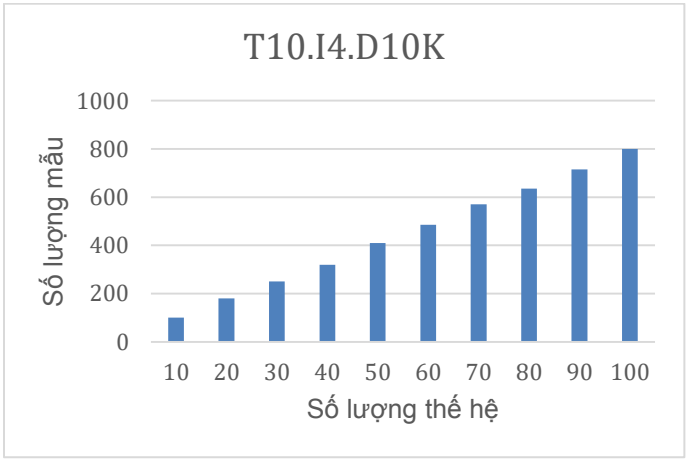
Để kiểm tra khả năng mở rộng của HUPE$_{UMU}$-GARM với số lượng thay đổi các giao dịch, thử nghiệm trên tập dữ liệu tổng hợp của IBM được sử dụng. Các ngưỡng *minUtil* được đặt thành 2%. Kết quả được trình bày trong Hình 9. HUPE$_{UMU}$-GARM hiển thị khả năng mở rộng tuyến tính với số lượng giao dịch từ 10 nghìn đến 100 nghìn. HUPE$_{UMU}$-GARM quy mô tốt hơn nhiều và nó hỗ trợ tập dữ liệu lớn.



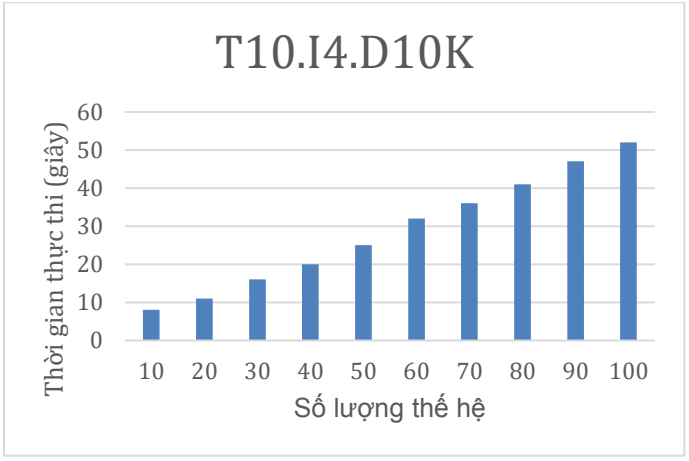
Hình 9. *Biểu diễn khả năng mở rộng với số lượng giao dịch (đvt: 1000)*

L. THUẬT TOÁN HUPE$_{WUMU}$-GARM

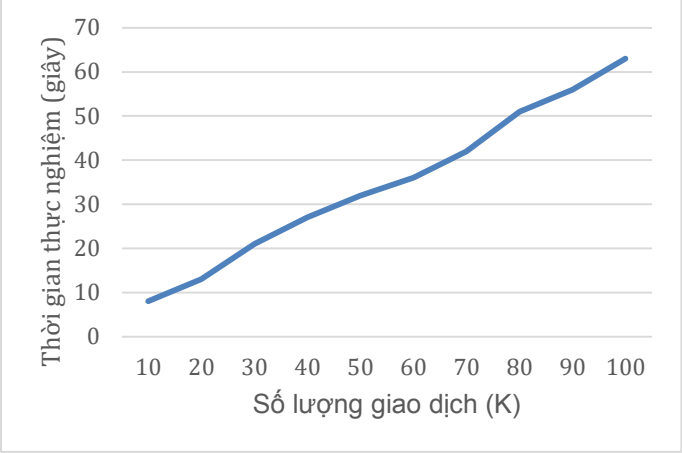
Thử nghiệm tiếp theo được thực hiện với HUPE$_{WUMU}$-GARM tiếp cận với các số lượng các thế hệ khác nhau bằng cách giữ cố định 100 mục. Kết quả thử nghiệm được minh họa qua Hình 10 và Hình 11 tương ứng. Giống với HUPE$_{UMU}$-GARM, thời gian thực thi của cách tiếp cận HUPE$_{WUMU}$-GARM được chứng minh là ít hơn đáng kể. Cùng với phương pháp tiếp cận HUPE$_{WUMU}$-GARM là một thử nghiệm trong việc tăng kích thước của giao dịch từ 10K đến 100K. Cách tiếp cận này cho thấy khả năng mở rộng tuyến tính và quy mô tốt hơn nhiều, được minh họa trong Hình 12.



Hình 10. *Biểu diễn số lượng mẫu tương ứng với số thế hệ.*



Hình 11. *Biểu diễn thời gian thực thi tương ứng số thế hệ*



Hình 12. *Thời gian thực thi ứng với số giao dịch (đvt: 1000)*

PHẦN KẾT LUẬN

Khai thác HUI là một vấn đề nghiên cứu rất quan trọng trong khai thác dữ liệu và khám phá tri thức. Các cách tiếp cận được đề cập trong bài viết này khai thác một GA; chúng có thể xử lý số lượng lớn các mục và giao dịch riêng biệt. Những phương pháp tiếp cận này là lựa chọn tốt nhất khi thời gian thực hiện và bộ nhớ yêu cầu là những vấn đề quan trọng. GAs đã được chứng minh là mạnh mẽ, những kỹ thuật tìm kiếm mục đích chung. Chúng có một vị trí trung tâm trong việc khai thác dữ liệu ứng dụng vì khả năng khám phá không gian tìm kiếm lớn một cách hiệu quả. Hai thách thức lớn trong khai thác độ hữu ích, đó là không gian tìm kiếm theo cấp số nhân và ngưỡng hữu ích tối thiểu phụ thuộc vào CSDL đã được nghiên cứu và một nỗ lực được thực hiện để giải quyết các vấn đề bằng cách đề xuất một cách tiếp cận dựa trên GA cho HUI từ CSDL. Những cách tiếp cận này hoạt động tốt trên các tập mục chứa các giá trị mục âm. Kết quả thử nghiệm cho thấy tiếp cận quy mô tốt và truy xuất HUI từ CSDL chứa giá trị mục âm một cách hiệu quả.

TÀI LIỆU THAM KHẢO

[1] Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," *Advances in knowledge discovery and data mining,* pp. 1-34, 1996.

[2] Agrawal, R., T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases,"*In Proceedings of ACM SIGMOD conference*, 1993.

[3] Agrawal, R., and R. Srikant, "Fast algorithms for mining association rules,"*In Proceedings of the 20th very large data bases (VLDB) conference*, San Francisco, CA, USA, 1994.

[4] Agrawal, R., and J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering 8(6),* pp. 962-969, 1996.

[5] Ahmed, C. F., S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *IEEE Transactions on Knowledge and Data Engineering,* p. 1708–1721, 2009.

[6] Ahmed, C. F., S. K. Tanbeer, and B.-S. Jeong, "Efficient mining of high utility patterns over data," in *In Software engineering, artificial intelligence, networking and*, Berlin, 2010.

[7] Chu, C.-J., V. S. Tseng, and T. Liang, "An efficient algorithm for mining high utility itemsets with negative item values in large databases," *Applied Mathematics and Computation 215(2),* pp. 767-778, 2009.

[8] Erwin, A., R. P. Gopalan, and N. R. Achuthan, "A bottom-up projection based algorithm for mining high utility itemsets,"*In Proceedings of the 2nd workshop on integrating ai and data mining (AIDM 2007)*, Darlinghurst, Australia, 2007.

[9] Li, H.-F, "MHUI-max: An efficient algorithm for discovering high-utility itemsets from data streams," *Journal of Information Science 37(5),* pp. 532-545, 2011.

[10] Liu, Y. W.-K. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," *Workshop on utility-based data mining (UBDM 2005),* pp. 90-99, 2005.

[11] Tseng, V. S., C.-J. Chu, and T. Liang, "Efficient mining of temporal high utility itemsets from data streams," in *Proceedings of the 2nd international workshop on utility-based data mining*, Philadelphia, PA, USA, 2006.

[12] Tseng, V. S., C.-W. Wu, B.-E. Shie, and P. S. Yu, "UP-growth: An efficient algorithm for high utility itemset mining," in *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, New York, 2010.

[13] Wu, C. W., B.-E. Shie, P. S. Yu, and V. S. Tseng, "Mining top-k high utility itemsets," in *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (ACM, KDD'12)*, Beijing, China, 2012.

[14] Yao, H., and H. J. Hamilton, "Mining itemset utilities from transaction databases," *Data and Knowledge Engineering 59,* pp. 603-626, 2006.

[15] Yin, J., Z. Zheng, and L. Cao, "USpan: An efficient algorithm for mining high utility sequential patterns," in *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (ACM, KDD'12)*, Beijing, China, 2012.

[16] Y. L. W.-K. C. A. N. Liu, "A two-phase algorithm for fast discovery of high utility itemsets," *Ho, T. B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS,* vol. 3515, pp. 689-695, 2005.

[17] Beasley, D., D. R. Bull, and R. R. Martin, "An overview of genetic algorithms," *University Computing 15(2),* pp. 58-69, 1993.

[18] Berson, A., and S. J. Smith, Data warehousing, data mining and OLAP, India: Tata McGraw-Hill, 2004.

[19] Yao, H., H. J. Hamilton, and C. J. Butz, "A foundational approach to mining itemset utilities from databases," in *Proceedings of the 3rd SIAM international conference on data mining*, Orlando, Florida, USA, 2004.

[20] Li, Y.-C., J.-S. Yeh, and C.-C. Chang, "Isolated items discarding strategy for discovering high utility itemsets," *Data and Knowledge Engineering 64,* pp. 198-217, 2008.

[21] Yen, S.-J., and Y.-S. Lee, "Mining high utility quantitative association rules," in *Data warehousing and knowledge discovery, Lecture Notes on Computer Science 4654*, Berlin, Heidelberg, 2007.

[22] Ahmed, C. F., S. K. Tanbeer, and B.-S. Jeong, "Mining high utility web access sequences in dynamic web log data," in *In 11th ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing*, 2010.

[23] Zhou, L., Y. Liu, J. Wang, and Y. Shi, "Utility-based web path traversal pattern mining," in *Paper preented at the 7th IEEE international conference on data mining workshops. October 28-31*, Omaha, Nebr., 2007.

[24] Ahmed, C. F., S. K. Tanbeer, and B. S. Jeong, "A framework for mining high utility web access sequences," *IETE Technical Review 28(1),* pp. 3-16, 2011.

[25] Thilagu, M., and R. Nadarajan, "Efficiently mining of effective web traversal patterns with average utility," in *Procedia Technology Journal: 2nd International Conference on Communication, Computing, and Security 6*, 2012.

[26] Liu, Y. W.-K. Liao, and A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Proceedings of the 9th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD 2005), ed. T. B. Ho, D. Cheung, and H. Liu. Lecture Notes in Artificial Intelligence 3518*, Berlin, Heidelberg, 2005.

[27] J. Holland, Adaptation in natural and artificial systems, Ann Arbor, MI, USA: University of Michigan Press, 1975.

[28] Premalatha, K., and A. M. Natarajan, "Genetic algorithm for document clustering with simultaneous and ranked mutation," *ournal of Modern Applied Science 3(2),* pp. 75-82, 2009.

[29] "IBM Quest Market-Basket Synthetic Data Generator," [Online]. Available: http://www.cs.loyola.edu/~cgiannel/assoc_gen.html.

HIGH UTILITY ITEMSETS MINING ALGORITHM BASED ON GENETIC WITH RANKED MUTATION

Pham Duc Thanh, Le Thi Minh Nguyen

ABSTRACT— Utility mining is the study of utility itemset mining from transactional database. It is a utility-based itemset mining approach to find itemsets that match user preferences. Recent research on mining high utility sets (HUIs) from databases faces two major challenges: Exponential search space and minimum utility threshold depends on the database. The search space is extremely large when the number of distinct items and the size of the database is very large. Data analysts must specify appropriate minimum utility thresholds for their mining tasks, even though they may not have the relevant knowledge of their database. Furthermore, a utility-mining algorithm supports only an itemset with positive item values. To avoid these problems, two approaches are presented to mine HUI containing negative item values from the transactional database: yes/no specified minimum utility threshold through a genetic algorithm with ranked mutation. According to our understanding, this is the first study in HUI mining with negative item values from transaction database using genetic algorithm. The experimental results show that the approaches described in this article achieve better performance in terms of scalability and efficiency.

Phạm Đức Thành.
Nhận học vị Thạc sĩ năm 2006 tại Đại học Quốc gia Thành phố Hồ Chí Minh; hiện đang là Giảng viên công tác tại khoa Công nghệ Thông tin Trường Đại học Ngoại ngữ-Tin học TP. Hồ Chí Minh, lĩnh vực nghiên cứu đang quan tâm là khai thác dữ liệu.

Lê Thị Minh Nguyện
Nhận học vị Thạc sĩ Khoa học máy tính tại Đại học Quốc gia Thành phố Hồ Chi Minh năm 2007. Hiện là giảng viên khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ-Tin học TP. Hồ Chí Minh. Lĩnh vực nghiên cứu đang quan tâm là khai thác dữ liệu.