ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

ĐỀ CƯƠNG ĐỀ TÀI LUẬN VĂN THẠC SĨ (PTII): 15TC

1. **Tên đề tài hoặc hướng NC** (*gồm cả tiếng Việt và tiếng Anh*):

Tiếng Việt: Khai thác tập phổ biến đóng trên cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số.

Tiếng Anh: Mining Frequent Weighted Closed Itemsets from Weighted Items Transaction Databases.

2. **Ngành và mã ngành đào tạo**: Ngành Khoa học máy tính, mã ngành: 06.48.01.01

3. **Họ tên học viên**: PHAN TẤN TÀI, khoá 10 – đợt 1

Địa chỉ email, điện thoại liên lạc của học viên: tantai.developer@gmail.com **- 0968 9696 77**

Người hướng dẫn: PGS.TS. VÕ ĐÌNH BẢY

Địa chỉ email, điện thoại liên lạc của người hướng dẫn: bayvodinh@gmail.com **- 0937 306 858**

4. Tổng quan tình hình nghiên cứu

4.1. Giới thiệu chung

Khai thác dữ liệu là quá trình tìm ra các tri thức mới từ một nguồn dữ liệu đã có. Khai thác tập phổ biến được đề xuất đầu tiên bởi Agrawal [1] vào năm 1993, là một bài toán quan trọng trong lĩnh vực khai thác dữ liệu. Khai thác tập phổ biến là quá trình tìm ra tất cả các tập hạng mục thoả mãn độ phổ biến tối thiểu cho trước. Từ đó đến nay, đã có nhiều thuật toán được phát triển để khai thác tập phổ biến như Apriori [2], Eclat [16], FP-Growth [8], N-list [5], Node-list [4], Nodeset [6], DiffNodeset [7].

Theo cách khai thác truyền thống, việc tìm tất cả các tập phổ biến rất tốn thời gian và tỏ ra không hiệu quả, vì vậy tác giả Pasquier [11] và Bastide [3] đã đưa ra một cách tiếp cận mới đó là tìm các tập phổ biến đóng (vì tập phổ biến đóng có số lượng nhỏ hơn tập phổ biến), đặc biệt đối với cơ sở dữ liệu đặc (mật độ trùng lắp các hạng mục giữa các dòng dữ liệu cao) hoặc ngưỡng hỗ trợ nhỏ thì khai thác tập phổ biến đóng tỏ ra hiệu quả hơn so với khai thác tập phổ biến. Một số thuật toán nghiên cứu về tập phổ biến đóng

như: A-Close [11] một mở rộng của Apriori, CHARM [18], CLOSET [12] thuật toán xây dựng trên phương pháp FP-Growth và CLOSET+ [15] cải tiến của thuật toán CLOSET.

Có nhiều phương pháp đề xuất cho khai thác tập phổ biến đóng và được chia thành bốn nhóm: tạo ứng viên và kiểm thử, chia để trị, lai ghép và lai ghép không trùng lắp.

Trong một số nghiên cứu còn đề cập đến một số lợi ích của các hạng mục (như giá cả của mặt hàng hay trọng số của các trang web) gắn liền với mỗi hạng mục. Ramkumar [13] đã đề xuất khai thác tập phổ biến có đánh trọng.

Bên cạnh đó, do số lượng tập phổ biến có đánh trọng khá lớn nên việc áp dụng tập phổ biến đóng có đánh trọng mang lại hiệu quả tính toán tốt hơn vì có ít tập phổ biến đóng có đánh trọng hơn so với tập phổ biến có đánh trọng.

Đã có các công trình nghiên cứu khai thác tập phổ biến đóng có đánh trọng như WIT-tree [14, 9, 10] nhưng phương pháp này có điểm hạn chế là tốn rất nhiều bộ nhớ để lưu trữ và xử lý. Nodeset là một cấu trúc khá tốt đã được áp dụng trên tập phổ biến và mang lại hiệu quả tính toán cao, đồng thời tiết kiệm được bộ nhớ vì vậy nghiên cứu tập trung tìm hiểu về phương pháp khai thác tập phổ biến đóng có đánh trọng dựa trên cấu trúc Nodeset.

4.2. Các nghiên cứu liên quan

a) Cơ sở dữ liệu giao dịch

Cơ sở dữ liệu giao dịch: Một cơ sở dữ liệu D được tạo thành từ tập giao dịch T = $\{t_1, t_2, \ldots, t_n\}$ và tập các hạng mục I = $\{i_1, i_2, \ldots, i_m\}$ được gọi là cơ sở dữ liệu giao dịch.

Cho hạng mục $X \subseteq I$, độ hỗ trợ của X trong D ký hiệu là sup(X) là số giao dịch trong D chứa X.

Tập phổ biến: là tập hợp các hạng mục thoả mãn độ phổ biến tối thiểu (minsupp - do người làm quy định), sup(S) ≥ minsupp thì S là tập phổ biến.

Tập phổ biến tối đại: là tập phổ biến và không có tập nào bao nó là tập phổ biến.

Tập phổ biến đóng: là tập phổ biến và không có tập cha nào có cùng độ hỗ trợ với nó.

b) Khai thác tập phổ biến

Thuật toán Apriori [2] được đề xuất bởi Agrawal và Srikant đề xuất vào năm 1994. Ý tưởng của thuật toán này là tìm các tập phổ biến một hạng mục, sau đó tạo ra các tập ứng viên có kích thước k-hạng mục từ tập phổ biến k-1 hạng mục, sau đó loại ra các tập ứng viên không phổ biến. Nguyên tắc cơ bản của thuật toán là nếu tập hiện tại không là tập phổ biến thì tập bao nó cũng không phổ biến. Thuật toán dễ hiểu, dễ cài đặt, tuy

nhiên thuật toán phải duyệt cơ sở dữ liệu nhiều lần và số lượng tập ứng viên rất lớn do đó không phù hợp cho những cơ sở dữ liệu lớn.

Trong thuật toán Eclat [16], mỗi hạng mục sẽ chứa danh sách các giao dịch chứa hạng mục đó, ta tìm độ hỗ trợ k-hạng mục bằng cách lấy giao hai danh sách giao dịch của (k-1) hạng mục con. Thuật toán này tốn khá nhiều bộ nhớ để lưu trữ danh sách các giao dịch cho các hạng mục.

Để khắc phục những hạn chế của thuật toán Apriori vào năm 2000, Han và các đồng sự đề xuất ra thuật toán FP-Growth [8], ý tưởng của thuật toán này là khai thác tập phổ biến mà không cần tạo ra các ứng viên. Thuật toán FP-Growth duyệt cơ sở dữ liệu để tìm ra các tập phổ biến một hạng mục, sắp xếp các hạng mục cơ sở dữ liệu theo độ phổ biến giảm dần. Sau đó tiến hành xây dựng cây FP và sử dụng cây FP tạo ra các cơ sở điều kiện từ đó khai thác tập phổ biến, ưu điểm của phương pháp này là không cần phải duyệt cơ sở nhiều lần như thuật toán Apriori, thuật toán thường có hiệu quả trên các cơ sở dữ liệu có mật độ trùng lắp dữ liệu cao. Tuy nhiên FP- Growth phải tốn nhiều bộ nhớ để lưu trữ cây FP và cài đặt khá phức tạp.

Cấu trúc Node-list [4] do Deng và Wang đề xuất vào năm 2010 và N-list [5] do Deng và các đồng sự đề xuất vào năm 2012 được sử dụng để khai thác tập phổ biến, cả hai cấu trúc này tạo ra node với thuộc tính pre-order và post-order. Dựa trên cấu trúc Node-list và N-list thuật toán PPV [4] và PrePost [5] được đưa ra để khai thác tập phổ biến. Tuy nhiên hai cấu trúc này tốn rất nhiều bộ nhớ bởi vì Node-list và N-list cần phải mã hóa node với thuộc tính pre-order và post-order. Hơn nữa cấu trúc Node-list và N-list không thích hợp cho việc nối các chuỗi Node-list và N-list tập hạng mục ngắn để tạo ra các chuỗi Node-list và N-list tập hạng mục dài.

Để khắc phục sự tốn kém bộ nhớ của cấu trúc N-list và Node-list, bằng cách loại bỏ đi một thành phần trong N-list và Node-list (loại bỏ pre-order hoặc post-order). Thuật toán FIN [6] với sự kết hợp giữa cấu trúc Nodeset và POC-tree có tốc độ thực thi nhanh hơn và tiết kiệm bộ nhớ hơn rất nhiều so với N-list và Node-list bởi vì chỉ sử dụng mỗi thuộc tính pre-order hoặc post-order để tìm kiếm các Nodeset.

DiffNodeset [7] một mở rộng từ Nodeset. Với cấu trúc DiffNodeset, thuật toán dFIN [7] được tạo ra. Ý tưởng của thuật toán này là dựa vào các Nodeset một hạng mục để tìm ra các DiffNodeset k hạng mục. Qua thực nghiệm thấy rằng, DiffNodeset có tốc độ thực thi và tốn bộ nhớ tương đương với Nodeset.

c) Khai thác tập phổ biến đóng

A-Close [11] áp dụng kỹ thuật tìm kiếm theo chiều rộng để tìm ra mô hình hình phổ biến đóng dựa trên các tập ứng viên. Thuật toán tỏ ra không hiệu quả trong trường hợp có nhiều ứng viên vì phải quét cơ sở dữ liệu nhiều lần.

CLOSET [12] là một dạng mở rộng của thuật toán FP-Growth. CLOSET xây dựng cây FP và duyệt đệ quy cây từ dưới lên để tìm ra tập phổ biến đóng. Tuy nhiên do kế thừa từ thuật toán FP-Growth nên thuật toán chỉ tối ưu trên những dữ liệu dày đặc, đồng thời tốn nhiều bộ nhớ để lưu trữ và duyệt cây FP.

CHARM [18] là một trong những thuật toán trong khai thác tập phổ biến đóng, ưu điểm của thuật toán này là không sinh ứng viên và dựa vào phương pháp chia để trị để tìm kiếm các tập phổ biến đóng và chỉ duyệt cơ sở dữ liệu một lần.

CLOSET+ [15] là một cải tiến của thuật toán CLOSET. Thuật toán CLOSET+ là một phương pháp lai. Trong khi thuật toán CLOSET chỉ quét theo chiều sâu của cây FP để tìm ra tập phổ biến đóng thì thuật toán CLOSET+ kết hợp tìm kiếm theo chiều sâu để tìm ra các hạng mục phổ biến cục bộ của một tiền tố nào đó, sau đó quét theo chiều ngang để tìm ra các tập phổ biến đóng, điều này giúp giảm không gian tìm kiếm. Qua thực nghiệm cho thấy thuật toán CLOSET+ có tốc độ thực thi nhanh hơn so với thuật toán CHARM.

d) Cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số

Một cơ sở dữ liệu D được tạo thành từ tập giao dịch T = $\{t_1, t_2, \ldots, t_n\}$, tập các hạng mục I = $\{i_1, i_2, \ldots, i_m\}$ và tập các giá trị trọng số W = $\{w_1, w_2, \ldots, w_m\}$ tương ứng với các hạng mục trong I được gọi là cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số.

Kết nối Galois: Đặt $\delta \subseteq I \times T$ có quan hệ nhị phân, trong đó I là tập các hạng mục và T là tập các giao dịch chứa trong cơ sở dữ liệu D. Đặt $X \subseteq I$ và $Y \subseteq T$. Đặt P(S) bao gồm tất cả các tập con của S. Hai ánh xạ giữa P(I) và P(T) được gọi là kết nối Galois như sau:

$t: P(I) \mapsto P(T), t(X) = \{y \in T \mid \forall x \in X, x\ \delta\ y\}$

$i: P(T) \mapsto P(I), i(Y) = \{x \in I \mid \forall y \in Y, x\ \delta\ y\}$

Trọng số của một giao dịch t_k được định nghĩa như sau:

$$tw(t_k) = \frac{\sum_{i_j \in t_k} w_j}{|t_k|}$$

Trọng số hỗ trợ của các hạng mục được định nghĩa như sau:

$$ws(X) = \frac{\sum_{t_k \in t(X)} tw(t_k)}{\sum_{t_k \in T} tw(t_k)}$$

Tập phổ biến có đánh trọng: là tập hợp các hạng mục có trọng số hỗ trợ thỏa mãn trọng số hỗ trợ tối thiểu, ws(X) ≥ minws thì X là tập phổ biến có đánh trọng.

Tập phổ biến đóng có đánh trọng: là tập phổ biến có đánh trọng và không tồn tại tập phổ biến có đánh trọng cha nào mà có trọng số hỗ trợ bằng nó.

e) Khai thác tập phổ biến có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số

Vào năm 1998, Ramkumar [13] đã đề xuất một mô hình khai thác luật kết hợp dựa trên cơ sở giao dịch có hạng mục được đánh trọng số, đồng thời giới thiệu thuật toán dựa trên Apriori [13] để khai thác tập phổ biến có đánh trọng.

Sử dụng cấu trúc WIT-tree [9, 10] để khai thác tập phổ biến có đánh trọng, bằng cách xây dựng WIT-tree như sau: tính trọng số hỗ trợ dựa vào Tidset, xây dựng các lớp tương đương một phần tử và sắp xếp theo thứ tự trọng số hỗ trợ giảm dần, từ các lớp tương đương một phần tử xây dựng các lớp tương đương k phần tử, sau đó dựa vào ngưỡng trọng số hỗ trợ tối thiểu (minws) và tính bao đóng của lớp tương đương để loại bỏ những nút không phổ biến. Cải tiến WIT-tree bằng cách dựa vào tính bao đóng của lớp tương đương để tiến hành cắt tỉa trong quá trình xây dựng cây.

Sử dụng Diffset để tính trọng số hỗ trợ là một cải tiến của cấu trúc WIT. Diffset tính toán trọng số hỗ trợ bằng cách tính toán sự khác biệt giữa hai Tidset trong cùng một lớp tương đương. Zaki và Gouda [17] đã chứng minh được rằng Diffset tính toán trọng số hỗ trợ và tiết kiệm bộ nhớ nhiều hơn so với Tidset.

f) Khai thác tập phổ biến đóng có đánh trọng trên cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số [14]

Dựa vào cấu trúc WIT-tree [14] để tìm ra các tập phổ biến có đánh trọng sau đó dựa vào tính chất của tập Tidset để loại bỏ những tập phổ biến có đánh trọng không phải là tập phổ biến đóng có đánh trọng. Hướng tiếp cận này phải tốn không gian lưu trữ các Tidset và thời gian tính toán các Tidset.

5. Tính khoa học và tính mới

Cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số (chẳng hạn như cơ sở dữ liệu giao dịch của hàng hoá với giá cả) khá gần gũi với thực tế. Vì vậy, khai thác tập phổ biến từ cơ sở giao dịch có hạng mục được đánh trọng số mang tính ý nghĩa thực tiễn trong việc hỗ trợ ra quyết định trong môi trường thực tế.

Cấu trúc Nodeset đã được chứng minh rất hiệu quả trong khai thác tập phổ biến trên cơ sở dữ liệu giao dịch. Tuy nhiên, hiện nay vẫn chưa có công trình nghiên cứu nào áp dụng cấu trúc Nodeset trên cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số để khai thác tập phổ biến đóng có đánh trọng. Vì vậy, nghiên cứu tìm cách áp dụng cấu trúc Nodeset để khai thác tập phổ biến đóng có đánh trọng trên cơ sở dữ liệu có hạng mục được đánh trọng.

6. Mục tiêu, đối tượng và phạm vi

6.1. Mục tiêu nghiên cứu

Mục tiêu tổng quan: nghiên cứu về cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số, tập phổ biến có đánh trọng và tập phổ biến đóng có đánh trọng. Áp dụng cấu trúc Nodeset vào để tìm ra các tập phổ biến đóng có đánh trọng.

Mục tiêu chi tiết:

- Tìm hiểu về cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số, tập phổ biến có đánh trọng, tập phổ biến đóng có đánh trọng.
- Tìm hiểu các phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số. Nhận xét ưu, khuyết điểm và đánh giá.
- Xây dựng phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số dựa trên cấu trúc Nodeset.

6.2. Đối tượng và phạm vi nghiên cứu

Đối tượng: trên dữ liệu lớn.

Phạm vi: khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số.

7. Nội dung, phương pháp

Nội dung 1: Tìm hiểu và cài đặt các phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số, nhận xét ưu khuyết điểm.

- Kết quả dự kiến: Báo cáo tổng quan về các phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số.
- Phương pháp thực hiện: Tìm hiểu, nghiên cứu thông qua các bài báo, công trình, tài liệu trên các trang có uy tính, giảng viên hướng dẫn cung cấp, … về các phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số.

Nội dung 2: Nghiên cứu và cài đặt phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số dựa trên cấu trúc Nodeset.

- Kết quả dự kiến: các lý thuyết về khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số dựa trên cấu trúc Nodeset.
- Phương pháp thực hiện: cài đặt phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số dựa trên cấu trúc Nodeset. Chạy thực nghiệm và so sánh với các phương pháp trước đó và rút ra kết luận.

8. Kế hoạch bố trí thời gian nghiên cứu

THÁNG	NỘI DUNG
Thứ 1, 2, 3	Nghiên cứu phương pháp, cài đặt phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số dựa trên cấu trúc Nodeset. Cài đặt các phương pháp khai thác tập phổ biến đóng có đánh trọng từ cơ sở dữ liệu giao dịch có hạng mục được đánh trọng số đã có. Viết bài báo khoa học.
Thứ 4,5	Chỉnh sửa và đăng bài báo khoa học. Viết luận văn.
Thứ 6	Chỉnh sửa và nộp luận văn.

9.Tài liệu tham khảo

[1] Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record, vol. 22, no. 2, pp. 207-216, 1993.

[2] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, vol. 1215, pp. 487-499, 1994.

[3] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Computational Logic CL, vol. 1861, pp. 972-986, Springer Berlin Heidelberg, 2000.

[4] Deng, Z., Wang, Z.: A new fast vertical method for mining frequent patterns. In: International Journal of Computational Intelligence Systems, vol. 3, no. 6, pp. 733-744, 2010.

[5] Deng, Z.H., Wang, Z.H., Jiang, J.J.: A new algorithm for fast mining frequent itemsets using N-lists. In: Science China Information Sciences, vol. 55, no. 9, pp. 2008-2030, 2012.

[6] Deng, Z.H., Lv, S.L.: Fast mining frequent itemsets using Nodesets. In: Expert Systems with Applications, vol. 41, no. 10, pp. 4505-4512, 2014.

[7] Deng, Z.H.: DiffNodesets: An efficient structure for fast mining frequent itemsets. In: Applied Soft Computing, vol. 41, pp. 214-223, 2016.

[8] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM Sigmod Record, vol. 29, no. 2, 2000.

[9] Le, B., Nguyen, H., Cao, T.A., Vo, B.: A novel algorithm for mining high utility itemsets. In: Intelligent Information and Database Systems, ACIIDS, First Asian Conference on IEEE, pp. 13-17, 2009.

[10] Le, B., Nguyen, H., Vo, B.: An efficient strategy for mining high utility itemsets. In: International Journal of Intelligent Information and Database Systems, vol. 5, no. 2, pp. 164-176, 2011.

[11] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: International Conference on Database Theory, vol. 1540, pp. 398-416, Springer Berlin Heidelberg, 1999.

[12] Pei, J., Han, J., Mao, R.: CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In: ACM SIGMOD workshop on research issues in data mining and knowledge discovery, vol. 4, no. 2, pp. 21-30, 2000.

[13] Ramkumar, G.D., Ranka, S., Tsur, S.: Weighted Association Rules: Model and Algorithm. In: Proc. ACM SIGKDD, 1998.

[14] Vo, B., Coenen, F., Le, B.: A new method for mining Frequent Weighted Itemsets based on WIT-trees. In: Expert Systems with Applications, vol. 40, no. 4, pp. 1256-1264, 2013.

[15] Wang, J., Han, J., Pei, J.: Closet+: Searching for the best strategies for mining frequent closed itemsets. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 236-245, 2003.

[16] Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New Algorithms for Fast Discovery of Association Rules. In: KDD, vol. 97, pp. 283-286, 1997.

[17] Zaki, M.J., Gouda, K.: Fast vertical mining using diffsets. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 326-335, ACM, 2003.

[18] Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. In: IEEE transactions on knowledge and data engineering, vol. 17, no. 4, pp. 462-478, 2005.

TP. HCM, ngày 04 tháng 11 năm 2016.

NGƯỜI HƯỚNG DẪN	**HỌC VIÊN KÝ TÊN**
(*Họ tên và chữ ký*)	(*Họ tên và chữ ký*)
PGS.TS. VÕ ĐÌNH BẢY	PHAN TẤN TÀI