|  |  |
| --- | --- |
| 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: Ứng dụng N-list trong khai thác mẫu tuần tự phổ biến

Tiếng Anh: Sequential pattern mining using N-lists.

**2**. **Ngành và mã ngành**: Ngành khoa học máy tính, mã ngành: 06.48.01.01

**3**. **Họ tên học viên** thực hiện đề tài, khóa- đợt học: BÙI VĂN BẰNG, khóa 10 - đợt 1

**Địa chỉ email, điện thoại liên lạc** của học viên:[vanbang0208@gmail.com](mailto:vanbang0208@gmail.com) - 0984219010

**Người hướng dẫn**: PGS.TS VÕ ĐÌNH BẢY

**Địa chỉ email, điện thoại liên lạc** của người hướng dẫn:bayvodinh@gmail.com - 0937306858

**4**. **Tổng quan tình hình NC** *trong nước và ngoài nước: mỗi nội dung đề cập cần có trích dẫn tài liệu, công trình; cần phân tích, đánh giá những kết quả NC mới có liên quan đến đề tài hoặc hướng NC; cần chỉ ra những gì còn tồn tại*:

**Bài toán thực tế**

Xem xét một cơ sở dữ liệu giao dịch của một cửa hàng bán sách, với một đối tượng đại diện cho một khách hàng, nhưng thuộc tính đại diện cho tác tác giả hay tên sách. Cơ sở dữ liệu ghi nhận những quyển sách được mỗi khách hàng mua trong một khoảng thời gian nhất định. Câu hỏi đặt ra là những quyển sách nào được khách hàng mua cùng nhau nhiều nhất trong một khoảng thời gian trên.

Một trang web nổi tiếng được nhiều người truy cập. Một đối tượng là một người dùng, thuộc tính là một trang của web. Bài toán là tìm ra chuỗi các trang của web được truy cập phổ biến.

**Bài toán khai thác mẫu tuần tự phổ biến được phát biểu đầy đủ như sau:**

Với I = { , , …, } là một tập m phần tử khác nhau, được gọi là những item. Một itemset là một tập hợp không có thứ tự và khác rỗng của những item (Xem như các item trong itemset được sắp sếp theo thứ tự tăng dần). Một chuỗi là một danh sách có thứ tự của các itemset. Một itemtset được ký hiêu ( … ) với là một item. Một itemset với k item được gọi là k-itemset. Một chuỗi S được ký hiệu ( → → ··· → ), với mỗi phần tử là một itemset. Số itemset trong chuỗi được gọi là kích thước của chuỗi, và tổng số item của chuỗi đươc gọi là độ dài của chuỗi. Một chuỗi có độ dài k được ký hiệu là k-sequence.

Một chuỗi S = ( → → ··· → ) được gọi là chuỗi con của chuỗi R = ( → → ··· → ), được ký hiệu S ⊆ R, nếu tồn tại một số nguyên → < ··· < sao cho ⊆ với tất cả .

Cho một cơ sở dữ liệu tuần tự D, với mỗi chuỗi có một id duy nhất, cho một chuỗi S = ( → → ··· → ), với độ phổ biến của S trong D được tính bằng số lần chuỗi D xuất hiên S. Với một ngưỡng phổ biến cho trước (ký hiệu là minsup), một chuỗi được gọi là tập phổ biến nếu nó có độ phổ biến thỏa mãn ngưỡng minsup.

Cho một cơ sở dữ liệu tuần tự D, và ngưỡng minsup, bài toán là tìm tất cả những chuỗi tuần tự phổ biến trong D.

**Thuật toán FP-Growth**

Thuật toán FP-Growth được đề xuất bởi Han, Pie và Yin vào năm 2000 để khai thác tập phổ biến của cơ sở dữ liệu giao dịch từ cấu trúc cây không phải sinh ra tập ứng viên, theo chiến lược chia để trị.

Cơ sở dữ liệu lớn ban đầu được đưa vào một dạng dữ liệu nén là FP-tree, cấu trúc cây này chỉ lưu nhưng mẫu phổ biến. Sau đó chia cở sở dữ liệu nén thành những cây FP-tree điều kiện tương ứng với những hạng mục phổ biến theo F-list. Quá trình khai thác là riêng biệt cho mỗi cơ sở dữ liệu điều kiện. Thuật toán FP-growth khai thác đệ quy và phát triển mẫu phổ biến cho đến khi cây FP điều kiện chỉ chứ một đường dẫn duy nhất.

Thuật toán này cho thấy ưu việt về tốc độ thực thi so với thuật toán được dùng phổ biến trước đó là Apriori.

**Thuật toán SPADE** []

Thuật toán SPADE gồm việc xác định chuỗi phổ biến 1 phần tử, chuỗi phổ biến 2 phần tử, chuyển vào dạng dữ liệu theo chiều dọc, và xác định những chuỗi phổ biến khác bằng BFS hay DFS search trong mỗi lớp.

Thuật toán SPADE qua thử nghiệm và thực tế cho thấy ưu việt hơn nhiều so với thuật toán GSP.

**Cấu trúc cây FUSP**

Để tăng cường hiệu quả của quá trình khai thác tập tuần tự phổ biến. Lin et đã đưa ra cấu trúc cây FUSP. Việc tạo ra cấu trúc cây này từ cơ sở dữ liệu ban đầu cũng tương tự như việc tao ra cây FP-tree. Nhưng khác nhau là liên kết giữ hai nút trên cây phần làm hai loại là ‘s’ và ‘i’. Liên kết ‘s’ là liên kết giữa hai sự kiện khác trong trong chuỗi. Liên kết ‘i’ là liên kết giữ nhưng item trong cùng một itemset.

Dự trên tư tưởng của thuật toán FP-growth, giải quyết bài toán khai thác tập phổ biến tuần tự không sinh ra tập ứng viên.

**Khai thác tập phổ biến dựa trên N-list (PrePost)**

Dựa trên cấu trúc cây PPC-tree. Cấu trúc cây PPC-tree gồm hai đặc điểm chính. Thứ nhất là một nút gốc với một tập nhưng nút con. Thứ hai mỗi nút con gồm năm thành phần: item-name, count, children-list, pre-order, post-order.

Cây PPC-tree khác cây FP-tree ở ba điểm chính là:

* PPC-tree không có header table để duy trì sự kết nối giữ các nút.
* Mỗi nút của cây PPC-tree có thêm thuộc tính là pre-order, post-order.
* FP-tree đươc sử dụng trong toàn bộ quá trình của thuật toán FP-growth trong khi đó PPC-tree chỉ để sinh ra Pre-Post của mỗi nút.

N-list là một tập có thứ tự của nhưng nút sinh ra từ cây PPC-tree. N-list có hai thuộc tính quan trong là N-list của itemset có độ dài k+1 được tạo thành từ việc join N-list của nhưng tập con có độ dài k. Và độ phổ biến của một itemset là count của nhưng nút trên N-list.

Prepost được kiểm nghiệm

**5**. **Tính khoa học** **và tính mới** của đề tài: *những hạn chế hiện thời, những vấn đề khoa học cần NC giải quyết và tính cấp thiết của nó, hướng giải quyết có gì mới, lợi ích khoa học và thực tiễn khi vấn đề được giải quyết*:

Bài toán tìm tập phổ biến tuần tự là một bài toán quan trọng của việc khai thác dữ liệu tuần tự.

Thuật toán PrePost dựa trên cây PPC-tree và cấu trúc N-list đã cho thấy hiệu quả trong việc tìm tập phổ biến trên cơ sở dữ liệu giao dịch.

Nhưng chưa có một phương pháp nào kết hợp cấu trúc N-list vào tìm tập phổ biến tuần tự.

Ứng dụng N-list trong thác mẫu phổ biến tuần tự. Đây là một cách tiếp cận mới cho bài toán tìm mẫu tuần tự phổ biến.

**6.** **Mục tiêu, đối tượng và phạm vi** NC đề tài *cần hướng tới và khả năng giải quyết: Nêu cụ thể, súc tích theo từng mục; không viết chung chung, lý giải dài dòng,..*:

6.1 Mục tiêu nghiên cứu

Mục tiêu tổng quan: Đưa cấu trúc N-list vào phương pháp khai thác tập phổ biến tuần tự dự trên cây FUSP.

Mục tiêu chi tiết:

Xây dựng nền tảng lý thuyết cho việc kết hợp N-list để khai thác tập phổ biến tuần tự trên cây FUSP.

Hiện thưc lý thuyết đã chứng minh.

Kiểm tra kết quả, so sách với các phương pháp khác SPADE, FUSP. Đưa ra nhận xét.

6.2 Đối tượng và pham vi nghiên cứu

* Giải quyết bài toán tìm tập phổ biến trên cơ sở dữ liệu tuần tự
* Cở sở dữ liệu chỉ chứa nhưng itemset đơn.

**7**. **Nội dung, phương pháp** dự định NC. *Trình bày theo từng bước, mỗi bước cần có: mục tiêu dự kiến kết quả***+***, phương pháp, bố trí NC cụ thể (case-study), phương án phối hợp với các đơn vị, phòng thí nghiệm,... kết quả dự kiến đạt được và sẽ công bố*:

+ Tìm hiểu về bài toán khai thác tập tuần tự phổ biến.

+ Đưa ra nhận xét hai cấu trúc N-list và FUSP với việc khai thác tập tuần tự phổ biến.

+ Kết hợp cấu trúc N-list và cây FUSP để khai thác tập phổ biến.

+ Hiện thực và đánh giá kết quả đạt được so với nhưng phương pháp trước đây như SPADE, FUSP.

**8**. **Kế hoạch bố trí thời gian NC** *phải kết hợp nhiệm vụ NC với học tập; làm xêmina, kế hoạch làm việc với người hướng dẫn, kế hoạch viết bản thảo luận văn,...* :

|  |  |
| --- | --- |
| Thời gian | Công việc |
|  | Đọc tài liệu tham khảo, đưa ra nhận xét mặt còn tồn tại và hướng giải quyết. |
|  | Nghiên cứu lý thuyết và đưa ra giải pháp làm nên tảng cho việc hiên thực. |
|  | Hiện thực lý thuyết đã nghiên cứu và chỉnh sửa cho phù hợp thực tế. |
|  | Viết báo cáo về kết quả thu được |
|  | Cải tiến phần hiện thực và lý thuyết sau khi được nhận xét từ phía thầy hướng dẫn và tạp chí cho hoàn thiện |

**9**.**Tài liệu tham khảo** *gồm những tài liệu có liên quan nhất với đề tài NC; ngoài những tài liệu kinh điển phải có những tài liệu cận đại (năm năm trở lại) liên quan đến đề tài.Viết theo mẫu***+***, và thuyết minh đề tài phải tham chiếu đến các tài liệu tham khảo này*.

[1] Agrawal R and Srikant R, “Mining Sequential Patterns”, in Int'l. Conf. Data Engineering

(ICDE 95), pp.3-14, 1995.

[2] M. Zaki, “SPADE: An Efficient Algorithm for Mining Frequent Sequences,” Machine Learning, vol. 40, pp. 31-60, 2001.

[3] C. W. Lin, T. P. Hong, Wen-Hsiang Lu and Wen-Yang Lin, “An Incremental FUSP-Tree Maintenance Algorithm,” The Eighth International Conference on Intelligent System Design and Application, pp.445-449,2008.

[4] Deng Z H, Wang Z H, Jiang J J. A new algorithm for fast mining frequent itemsets using N-lists. Sci China Inf Sci, 2012, 55: 2008–2030, doi: 10.1007/s11432-012-4638-z

*TP. HCM, ngày tháng năm 201..*

**NGƯỜI HƯỚNG DẪN HỌC VIÊN KÝ TÊN**

(*Họ tên và chữ ký*) (*Họ tên và chữ ký*)

…………………………………. ………………………………….

***Một số lưu ý*:** *Văn phong* nên diễn đạt cô đọng, hình ảnh, sơ đồ, biểu bảng,.. rõ ràng; phân tích ý nghĩa nghiên cứu đề tài và về kết quả dự kiến một cách thuyết phục. *Dự kiến kết quả*, có thể là: Đề xuất quan điểm; Đưa ra báo cáo hiện trạng; Đề xuất các biện pháp về tổ chức, quản lý; Tạo lập được mô hình, sản phẩm, công nghệ mẫu,v.v.. *Tài liệu tham khảo*: viết theo dạng mẫu, là sách hay bài báo, ví dụ:

+ *với sách tham khảo*:

[1]. Nguyen P. K. (2005), Flood Simulation in the Vietnam Mekong Delta. *Proceedings of the 3rd International Symposium on Flood Defense*. Taylor & Francis Publ., pp.611-616. Netherlands.

+ *với bài báo khoa học*:

[2]. Hai Bang T., Ngoc Thanh N., Phi Khu N. (2011), “*Fuzzy Ontology Building and Integration for Fuzzy Inference Systems in Weather Forecast Domain*”. LNAI 6592, the 3rd Intl’ Conference, ACIIDS 2011, Korea. Springer, ISSN 0302-9743, ISBN 978-3-642-20042-7.