THÔNG TIN VỀ LUẬN ÁN

Tên đề tài luận án:	**Khai thác mẫu tuần tự phổ biến dựa trên ràng buộc**
Chuyên ngành:	Khoa học máy tính
Mã số ngành:	62 48 01 01
Họ tên nghiên cứu sinh:	VĂN THỊ THIÊN TRANG
Người hướng dẫn khoa học:	GS. TS. Lê Hoài Bắc
Cơ sở đào tạo:	Đại học Công nghệ Thông tin – Đại học Quốc gia Tp.HCM

1. TÓM TẮT NỘI DUNG LUẬN ÁN

Luận án trình bày tổng quan, tìm hiểu cơ sở lý thuyết về khai thác mẫu tuần tự phổ biến dựa trên các ràng buộc. Trong đó, đi sâu vào nghiên cứu vấn đề khai thác mẫu tuần tự với ràng buộc Itemset, và ứng dụng tập mẫu thỏa ràng buộc trong khai thác luật tuần tự có ràng buộc. Bên cạnh đó, luận án nghiên cứu lĩnh vực ứng dụng cụ thể của mẫu tuần tự là khai thác sử dụng web theo ràng buộc chuỗi con. Luận án đã hoàn thành được mục tiêu là đề xuất các phương pháp khai thác hiệu quả cho các bài toán đặt ra sao cho có thể tìm được tập mẫu thỏa ràng buộc một cách chính xác, rút ngắn thời gian khai thác và giảm bộ nhớ sử dụng.

Kết quả của luận án được tổng hợp từ 4 bài báo đã công bố trên các tạp chí: Knowledge and Information Systems (1 bài), Vietnam Journal Computer Science (Springer, 1 bài), Applied Intelligence (2 bài), và được thể hiện ở 3 bài toán chính sau:

Bài toán 1: khai thác mẫu tuần tự dựa trên ràng buộc Itemset là đi tìm tất cả các mẫu phổ biến trong cơ sở dữ liệu chuỗi, mà có chứa bất kỳ Itemset nào trong tập Itemset ràng buộc do người dùng đưa ra nhằm giảm bớt số lượng mẫu dư thừa, thực hiện khai thác có tập trung vào những itemset mà người dùng quan tâm, rút ngắn thời gian khai thác và giảm bộ nhớ sử dụng.

Bài toán 2: từ kết quả nghiên cứu của *bài toán 1*, luận án mở rộng mục tiêu phát triển thuật toán khai thác luật tuần tự có ràng buộc Itemset ở vế trái luật bằng cách tận dụng tập mẫu thỏa ràng buộc Itemset khai thác được cho quá trình sinh luật.

Bài toán 3: khai thác mẫu truy cập web dựa trên ràng buộc chuỗi con là đi tìm tất cả các mẫu phổ biến trong cơ sở dữ liệu chuỗi truy cập web mà có chứa bất kỳ chuỗi nào của tập chuỗi ràng buộc do người dùng chỉ ra dưới dạng chuỗi con.

2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN

Luận án này đưa ra các kết quả mới sau:

Đề xuất thuật toán khai thác mẫu tuần tự dựa trên ràng buộc Itemset: thuật toán *MSPIC-DBV*. Thuật toán mở rộng và phát triển cách tổ chức dữ liệu biểu diễn dọc, đề xuất cấu trúc *DBVP* làm đại diện biểu diễn lại cơ sở dữ liệu theo chiều dọc nhờ vậy chỉ duyệt cơ sở dữ liệu một lần. Bằng cách sử dụng cấu trúc cây tiền tố kết hợp *DBVP* để lưu không gian tìm kiếm, thuật toán đưa ra kỹ thuật tỉa không gian con theo tiền tố và kỹ thuật kiểm tra ràng

buộc theo tiền tố có thể bỏ qua việc kiểm tra ràng buộc cho một số lượng lớn các mẫu ứng viên.

Đề xuất thuật toán khai thác luật tuần tự với ràng buộc Itemset ở vế trái của luật gồm bộ ba thuật toán *MSRIC-B*, *MSRIC-R* và *MSRIC-P*. Trong đó, *MSRIC-B* là phương pháp cơ sở đơn giản đưa ràng buộc vào sau quá trình khai thác, hai thuật toán còn lại đưa vào trong quá trình khai thác. *MSRIC-R* đưa ở giai đoạn sinh luật, còn *MSRIC-P* đưa ở giai đoạn tìm mẫu, tận dụng kết quả của thuật toán *MSPIC-DBV*. *MSRIC-P* là thuật toán đóng góp chính, hiệu quả hơn hai thuật toán còn lại.

Đề xuất hai thuật toán khai thác mẫu truy cập web dựa trên ràng buộc chuỗi con gồm *MWAPC* và *EMWAPC*. Trong đó, *EMWAPC* là thuật toán đóng góp chính, cải tiến của *MWAPC*. *EMWAPC* sử dụng cấu trúc dữ liệu và các kỹ thuật tương tự phương pháp khai thác mẫu với ràng buộc Itemset. Tuy nhiên, dựa vào đặc điểm của mẫu truy cập web, thuật toán thực hiện tỉa nhanh không gian tìm kiếm ngay từ đầu và giảm thiểu việc kiểm tra ràng buộc dựa vào đặc điểm của ràng buộc chuỗi con.

3. CÁC ỨNG DỤNG/KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU

Trong tương lai, chúng tôi sẽ mở rộng nghiên cứu theo các hướng sau:

Tiếp tục phát triển các chiến lược tỉa không gian tìm kiếm hiệu quả cho bài toán khai thác mẫu tuần tự có ràng buộc để các thuật toán đạt tốc độ và bộ nhớ tối ưu hơn.

Nghiên cứu khai thác mẫu tuần tự có ràng buộc trên cơ sở dữ liệu phân tán, nhằm tìm cách xử lý hiệu quả cho các cơ sở dữ liệu cực lớn với chuỗi dữ liệu dài. Trong lĩnh vực khai thác thói quen sử dụng web, có thể áp dụng khai thác phân tán để khai thác web log bị phân tán trên nhiều server.

Nghiên cứu áp dụng các kỹ thuật đề xuất cho vấn đề khai thác mẫu tuần tự với các loại ràng buộc khác như: ràng buộc trong việc kết hợp các sự kiện của mẫu, ràng buộc thời gian.

CÁN BỘ HƯỚNG DẪN	NGHIÊN CỨU SINH
Lê Hoài Bắc	**Văn Thị Thiên Trang**