



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP

Tên đề tài:

KHAI THÁC LUẬT CÓ THỨ TỰ BẢN PHẦN TRONG CƠ SỞ DỮ LIỆU CHUỖI
MINING PARTIALLY-ORDERED EPISODE RULES IN AN EVENT SEQUENCE

Họ tên, chức danh GVHD: GS.TS.Lê Hoài Bắc

Sinh viên: Lê Hồng Quang - 18127190

Lê Thành Nam - 18127158

Thể loại KLTN: Nghiên cứu

Thời gian thực hiện: (từ 1/2022 đến ngày 7/2022)

Nội dung KLTN:

1. Giới thiệu đề tài:

Khai thác luật phổ biến(Frequent episode rule mining) là một công đoạn quan trọng trong ngành khai thác dữ liệu phục vụ cho mục đích phân tích chuỗi các sự kiện hoặc ký tự và từ các luật được tạo nên, dữ liệu sẽ được thấu hiểu và khám phá một cách dễ dàng.

Vấn đề cốt lõi của khai thác luật phổ biến là các sự kiện phải được sắp xếp theo thứ tự chặt chẽ. Chính vì vậy, sẽ có các luật tương tự nhau lại được xem là khác nhau nhưng trên thực tế lại miêu tả cùng một trường hợp. Để có thể tìm ra các luật mang tính tổng quát hơn và có khả năng thay thế phần lớn luật, khóa luận này nghiên cứu và giới thiệu một cách tiếp cận mới để hình thành nên các luật – thuật toán Partially-Ordered Episode Rules Miner (POERM).

2. Phạm vi khóa luận:

Khóa luận nghiên cứu các vấn đề cơ bản trong lĩnh vực khai thác dữ liệu, cùng với đó là thuật toán POERM giúp giải quyết vấn đề khai thác các tập luật trong cơ sở dữ liệu. Đồng thời, thuật toán POERM sẽ được cài đặt và cải tiến bằng ngôn ngữ lập trình Python.

3. Mục tiêu khóa luận:

Mục tiêu của khóa luận là tìm hiểu về thuật toán khai thác luật có thứ tự trong cơ sở dữ liệu chuỗi - một phương pháp mới giúp khai thác luật trong cơ sở dữ liệu chuỗi có độ hiệu quả cao hơn, tốc độ nhanh hơn, ít tốn bộ nhớ hơn so với các phương pháp khai thác luật đã biết. Sau đó sẽ đề xuất ý tưởng để cải tiến hiệu suất của thuật toán này, đồng thời đánh giá mức ảnh hưởng của các tham số đến thời gian thực thi và lượng bộ nhớ tiêu tốn qua các bộ dữ liệu thực tế.

4. Phương pháp tiếp cận:

Ý tưởng của thuật toán là sẽ bắt đầu tìm thành phần của luật $X \rightarrow Y$ theo thứ tự từ trái sang phải. Thuật toán POERM sử dụng quy tắc cắt giảm để xem xét tính hợp lệ của các luật giảm không gian tìm kiếm luật gốc (với m event set sẽ sinh ra $(2^m - 1) \times (2^m - 1)$ luật). Thuật toán xây dựng một bộ POER chứa các luật $X \rightarrow Y$ hợp lệ. Đầu tiên, thuật toán sẽ tìm ra các chuỗi 1-sự kiện thỏa định nghĩa cắt giảm và có khả năng là tiền tố X của luật $X \rightarrow Y$ đồng thời lưu lại các thời điểm X diễn ra. Sau đó thuật toán tiến hành mở rộng chuỗi X thành các chuỗi 2-sự kiện từ chuỗi X 1-sự kiện cũ và cứ tiếp tục mở

rộng đến khi không mở thêm được nữa và cũng lưu lại thời điểm của các sự kiện này. Tiếp đến, thuật toán tìm các hậu tố Y có thể là hậu tố của luật $X \rightarrow Y$ với tiền tố X được lưu trước đó cùng các thời điểm của chúng. Sau đó thuật toán sẽ cắt giảm hậu tố Y theo quy tắc cắt giảm. Tiếp đến, hậu tố Y cũng sẽ mở rộng như tiền tố X. Cuối cùng, thuật toán ghép tiền tố và hậu tố thành $X \rightarrow Y$ và cắt giảm theo quy tắc cắt giảm để cho các bộ POER hợp lệ. Thời điểm của các bước trên đều được lưu lại. Bên cạnh đó, quy trình MiningXEventSet được sử dụng để hỗ trợ việc cắt giảm các chuỗi tiền tố X cũng như mở rộng chúng.

5. Các mốc thời gian nghiên cứu:

STT	Nội dung	Mốc thời gian	Sinh Viên thực hiện
1	Tìm hiểu về bài toán khai thác dữ liệu và các thuật toán khai thác tập luật cơ bản.	14/1 – 31/1	Quang + Nam
2	Tiến hành nghiên cứu tài liệu về đề tài mà GVHD đã cung cấp	1/2– 14/2/	Quang + Nam
3	Nghiên cứu quy trình MiningXEvent phụ trợ cho thuật toán POERM.	16/2– 21/2	Quang
4	Nghiên cứu thuật toán POERM, các tham số cũng như các cấu trúc dữ liệu được sử dụng trong thuật toán.	16/02 - 21/02	Nam

5	Thực thi hàm MiningXEvent trên giấy bằng bộ dữ liệu đơn giản từ (ii)	22/2– 1/3	Nam
6	Thực thi thuật toán POERM trên giấy bằng bộ dữ liệu đơn giản từ (ii)	22/2– 1/3	Quang
7	Thực thi thuật toán POERM trên chương trình chạy bằng ngôn ngữ Java được cung cấp bởi (ii)	1/3 – 5/3	Quang + Nam
8	Viết chương trình biên dịch thuật toán POERM bằng ngôn ngữ Python	6/3 – 31/3	Nam
9	Phát triển và nghiên cứu hướng cải tiến cho thuật toán	1/4 – 16/4	Quang
10	Thực thi thuật toán trên các bộ dữ liệu thực tế từ (ii).	17/4 – 30/4	Quang + Nam
11	So sánh hai phiên bản thuật toán và khảo sát ảnh hưởng các tham số đối với thuật toán trên các bộ dữ liệu thực tế từ (ii).	1/5 – 31/5	Quang + Nam

12	Chỉnh sửa, bổ sung khóa luận tốt nghiệp và thiết kế slide trình bày.	1/6 – 30/6	Quang + Nam
13	Cập nhật đề cương và hoàn tất cuốn luận cho khóa luận tốt nghiệp.	1/7 – 27/7	Quang + Nam

6. Kết quả dự kiến của đề tài:


Dự kiến sẽ cài đặt thành công thuật toán POERM thành công. Bên cạnh đó, sử dụng thuật toán POERM để đánh giá thực nghiệm qua các bộ dữ liệu thực tế và sẽ thu được hiệu suất cao hơn so với các thuật toán khai thác tập luật truyền thống về thời gian chạy cũng như sự tiêu tốn bộ nhớ. Ngoài ra khảo sát được mức ảnh hưởng của các tham số đến thời gian thực thi và lượng bộ nhớ tiêu tốn của thuật toán, đồng thời phiên bản thuật toán POERM cải tiến sẽ thu được hiệu suất tốt hơn so với phiên bản đã nghiên cứu.


Ý kiến của giảng viên hướng dẫn


TP. HCM, 03/08/2022

Chữ ký của giảng viên hướng dẫn

Chữ ký (các) sinh viên


Lê Hải Thủy


Lê Hồng Quang


Lê Thành Nam