TRANG THÔNG TIN LUẬN ÁN TIẾN SĨ

Tên luận án: **Khai phá dữ liệu tuần tự để dự đoán hành vi truy cập Web**

Chuyên ngành: **Hệ thống thông tin**

Mã số: **9.48.01.04**

Họ và tên NCS: **Nguyễn Thôn Dã**

Người hướng dẫn khoa học: **TS. Tân Hạnh và TS. Phạm Hoàng Duy**

Cơ sở đào tạo: **Học viện Công nghệ Bưu chính Viễn thông**

NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN

- Đóng góp thứ nhất: Đề xuất một giải pháp để thiết kế và xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web. Luận án sử dụng 4 tập dữ liệu được thu thập từ các Website periwinklelecottages.com, palmviewsanibel.com, devqa.robotec.co.il và inees.org. Bài toán đặt ra là làm cách nào để tạo ra một cơ sở dữ liệu tuần tự từ tập hợp các tập tin Weblog. Ý tưởng chính của giải pháp là: Trong tập dữ liệu Weblog tìm một mảng chứa các IP khác nhau và một mảng chứa các liên kết khác nhau. Với mỗi các IP khác nhau có một nhóm các liên kết được truy cập theo thứ tự thời gian. Những nhóm này sẽ là các chuỗi dữ liệu tuần tự của cơ sở dữ liệu tuần tự cần tạo. Hơn nữa, bằng cách phân tích các đặc trưng của dữ liệu Weblog, luận án trình bày làm cách nào để chuyển đổi dữ liệu Weblog thành cơ sở dữ liệu tuần tự bằng một giải thuật tính toán song song và không song song.

- Đóng góp thứ hai: Đề xuất một giải pháp để làm giảm thời gian dự đoán cho dự đoán truy cập Web. Luận án sử dụng năm cơ sở dữ liệu tuần tự để thực hiện. Các cơ sở dữ liệu sử dụng gồm hai cơ sở dữ liệu được tạo ra từ các tập dữ liệu Weblog (thu thập từ các Website (palmviewsanibel.com và inees.org) và ba cơ sở dữ liệu click-stream là KOSARAK, FIFA và MSNBC. Bài toán được đặt ra là làm cách nào để dự đoán một trang kế tiếp theo sao một chuỗi S cho trước trong một cơ sở dữ liệu tuần tự SDB cho trước với một thời gian dự đoán tốt. Để giải quyết vấn đề này, luận án đề xuất năm bước chính: (i) Nhập vào cơ sở SDB và chuỗi tuần tự S; (ii) Loại bỏ các chuỗi tuần tự trong SDB mà không chứa các phần tử của chuỗi tuần tự S. Với các chuỗi tuần tự mà chứa các phần tử thuộc S, loại bỏ các chuỗi tuần tự trong SDB mà chỉ chứa duy nhất các phần tử của chuỗi tuần tự S ở vị trí cuối cùng. Giải pháp này sẽ làm giảm kích cỡ của cơ sở dữ liệu tuần tự gốc. Dựa

vào giải pháp này, thời gian dự đoán trên cơ sở dữ liệu tuần tự thu gọn nhanh hơn thời gian dự đoán của cơ sở dữ liệu gốc (chưa thu gọn). Đối với các tập dữ liệu được thu thập từ các tập tin Weblog, kết quả thử nghiệm trên tập dữ liệu palmviewsanibel.com cho thấy rằng thời gian dự đoán của mô hình đề xuất nhanh hơn 2.7 lần so với thời gian dự đoán của mô hình thông thường mà vẫn đảm bảo độ chính xác. Tương tự, kết quả thử nghiệm trên tập dữ liệu inees.org chỉ ra rằng thời gian dự đoán của mô hình đề xuất nhanh gần 2 lần so với thời gian dự đoán của mô hình thông thường. Với các tập dữ liệu click-stream, kết quả thử nghiệm trên FIFA, KOSARAK, MSNBC cho thấy rằng thời gian dự đoán của mô hình đề xuất nhanh lần lượt 3 lần, 30 lần, và 103 lần so với thời gian dự đoán của mô hình thông thường mà vẫn đảm bảo độ chính xác. Như vậy thực thi dự đoán trên các tập dữ liệu click-stream hiệu quả hơn nhiều so với thực thi dự đoán trên các tập dữ liệu thu thập từ các tập tin Weblog.

- Đóng góp thứ ba: Đề xuất một giải pháp để tăng độ chính xác cho dự đoán truy cập Web. Luận án sử dụng 3 cơ sở dữ liệu tuần tự để thực hiện giải pháp này. Các cơ sở dữ liệu tuần tự được thu thập từ các tập dữ liệu click-stream: KOSARAK, FIFA và MSNBC. Dựa trên đặc tính của PageRank và giải thuật CPT+, bài toán được đặt ra là làm cách nào để dự đoán một trang kế tiếp theo sau một chuỗi tuần tự cho trước trong một cơ sở dữ liệu tuần tự cho trước với một giải pháp tốt về độ chính xác. Luận án đề xuất 5 bước quan trọng của giải quyết vấn đề này: (i) Nhập vào một cơ sở dữ liệu tuần tự, (ii) Chuyển đổi các liên kết thành các nút của một cơ sở dữ liệu đồ thị, (iii) Tính toán PageRank cho từng nút, (iv) Tính toán trung bình PageRank cho mỗi chuỗi dữ liệu tuần tự, (v) Loại bỏ các chuỗi tuần tự có trung bình Page thấp sao cho độ chính xác của cơ sở dữ liệu thu gọn vẫn cao hơn độ chính xác của cơ sở dữ liệu tuần tự gốc (chưa thu gọn). Kết quả thử nghiệm cho thấy rằng giải pháp đề xuất cho độ chính xác cao hơn độ chính xác của tiếp cận thông thường khi thực hiện trên các tập dữ liệu khác nhau. Cụ thể là, trên cơ sở dữ liệu tuần tự MSNBC, khi giảm kích cỡ của cơ sở dữ liệu gốc (loại bỏ các chuỗi tuần tự có trung bình PageRank thấp) đến 50%, độ chính xác đã tăng lên đến 25%; trên cơ sở dữ liệu FIFA, khi giảm kích cỡ của cơ sở dữ liệu tuần tự gốc đến 15%, độ chính xác tăng đến 0.013%; trên cơ sở dữ liệu KOSARAK, khi giảm kích cỡ cơ sở dữ liệu tuần tự đến 15% thì độ chính xác tăng lên đến 0.027%.

- Đóng góp thứ tư: Đề xuất một mô hình kết hợp giữa tăng độ chính xác và giảm thời gian dự đoán. Luận án sử dụng cơ sở dữ liệu tuần tự KOSARAK, là cơ sở dữ liệu lớn nhất được

dùng trong luận án, đề làm dữ liệu đầu vào cho giải pháp này. Bằng phương pháp kiểm tra chéo K-Folder-Validation (với K = 10), cơ sở dữ liệu tuần tự KOSARAK đã được chia thành thành 10 phần ngẫu nhiên. Mỗi phần gồm 90% dữ liệu dùng cho huấn luyện và 10% còn lại dùng cho kiểm thử (dự đoán). Kết quả thử nghiệm chỉ ra rằng khi giảm kích cỡ cơ sở dữ liệu tuần tự gốc đến 34% (dùng giải pháp được trình bày trong phần Đóng góp thứ ba), độ chính xác trung bình của giải pháp đề xuất vẫn tốt hơn độ chính xác của tiếp cận thông thường. Tiếp theo, dùng 66% kích cỡ của sơ sở dữ liệu gốc (đã loại bỏ các dữ liệu thừa bằng giải thuật PageRank) để dự đoán bởi giải pháp được trình bày trong Đóng góp thứ hai, kết quả thực nghiệm chứng minh rằng độ chính xác trung bình đã tăng 0.0621% và thời gian dự đoán nhanh hơn xấp xỉ 80 lần so với tiếp cận thông thường.

CÁC ỨNG DỤNG, KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HOẶC NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU

- Nghiên cứu sâu hơn về dự đoán chuỗi dữ liệu tuần tự để phát triển những giải thuật mới nhằm giải quyết tốt hơn các vấn đề liên quan đến dự đoán truy cập Web.
- Các thử thách quan trọng về Big Data bao gồm thu thập dữ liệu, lưu trữ dữ liệu, phân tích dự liệu, tìm kiếm, chia sẻ, chuyển đổi, trực quan hóa dữ liệu …Do vậy, Big Data thường chứa dữ liệu có kích thước vượt quá sức chứa của phần mềm thông thường. Vì lý do này, dự đoán trên Big Data vẫn còn là một vấn đề mở và đưa ra những vấn đề lớn cần giải quyết. Trong tương lai, hướng phát triển của luận án là làm cách nào để giải quyết hiệu quả vấn đề dự đoán dữ liệu tuần tự trên Big Data về mặt thời gian và độ chính xác.

Xác nhận của người hướng dẫn khoa học 1 **Nghiên cứu sinh**

TS. Tân Hạnh **Nguyễn Thôn Dã**

Xác nhận của người hướng dẫn khoa học 2

TS. Phạm Hoàng Duy