HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Thôn Dã

KHAI PHÁ DỮ LIỆU TUẦN TỰ ĐỂ DỰ ĐOÁN HÀNH VI TRUY CẬP WEB

LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - Năm 2020

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Thôn Dã

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

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

TS. TÂN HẠNH

TS. PHẠM HOÀNG DUY

Hà Nội – Năm 2020

LỜI CAM ĐOAN

Tôi xin cam đoan luận án tiến sĩ ***Khai phá dữ liệu tuần tự để dự đoán hành vi truy cập Web*** là công trình nghiên cứu khoa học độc lập của riêng tôi. Các số liệu trong luận án có nguồn gốc xuất xứ rõ ràng. Các kết quả nghiên cứu trong luận án do tôi tự tìm hiểu, phân tích một cách trung thực, nghiêm túc, khách quan và chưa từng được công bố trong bất kỳ công trình nào khác.

Tác giả

Nguyễn Thôn Dã

LỜI CÁM ƠN

Tôi xin chân thành gửi lời cám ơn đến Ban lãnh đạo Học viện Công nghệ Bưu chính Viễn thông, Đào tạo Sau Đại học và tập thể thầy cô Khoa Công nghệ Thông tin đã có nhiều hỗ trợ cho tôi hoàn thành nhiệm vụ nghiên cứu được giao.

Tôi cũng gửi lời biết ơn đến hai cán bộ hướng dẫn luận án cho tôi là Thầy TS. Tân Hạnh và Thầy TS. Phạm Hoàng Duy (công tác tại Học viện Công nghệ Bưu chính Viễn thông), những người thầy với những kinh nghiệm và kiến thức chuyên môn cao đã tận tình hướng dẫn, chỉ bảo cho tôi để tôi có thể hoàn thành luận án này.

Tôi cũng rất cám ơn Ban Giám Hiệu trường Đại học Kinh tế - Luật, ĐHQG-HCM, nơi tôi đang công tác, đặc biệt là lãnh đạo Khoa Hệ thống thông tin của trường đã giới thiệu và tạo điều kiện cho tôi thực hiện luận án này.

Rất trân trọng và cám ơn các nhà nghiên cứu, các thầy cô, các đồng nghiệp đã có những góp ý hữu ích, phản biện khách quan và mang tính xây dựng để tôi không ngừng hoàn thiện luận án này.

Tôi cũng vô cùng biết ơn bố mẹ tôi, những người đã có công sinh thành và dưỡng dục, luôn động viên và giúp đỡ tôi trong suốt thời gian nghiên cứu và thực hiện luận án.

MỤC LỤC

DANH MỤC CÁC CHỮ VIẾT TẮT

TT	Chữ viết tắt	Nghĩa tiếng Anh	Nghĩa tiếng Việt
1.	AKOM	All-K-Order-Markov	Mô hình Markov thứ tự K
2.	ARM	Association Rule Mining	Khai phá luật tuần tự
3.	CLF	Common Log Format	Định dạng tập tin văn bản chuẩn được sử dụng bởi các máy chủ khi tạo ra các tập tin nhật ký máy chủ
4.	CPT	Compact Prediction Tree	Cây dự đoán nén
5.	DG	Dependency Graph	Đồ thị Phụ thuộc
6.	HITS	Hyperlink-Induced Topic Search	Giải thuật Tìm kiếm Chủ đề theo Siêu liên kết
7.	IPM	Integrated Prediction Model	Mô hình Dự đoán Tích hợp
8.	INR	Improved Noise Reduction	Chiến lược giảm nhiễu thông tin cải tiến
9.	IoT	Internet of Things	Mạng lưới vạn vật kết nối Internet
10.	LZ78	Abraham **L**empel & Jacob **Z**iv (19**78)**	Giải thuật LZ78: Giải thuật nén dữ liệu không mất thông tin được đề xuất 1978
11.	PPM	Prediction by Partial Matching	Giải thuật Dự đoán nén dữ liệu bằng So khớp Một phần
12.	SCM	Spare Count Matrix	Ma trận đếm thưa
13.	SPM	Sequential Pattern Mining	Khai phá mẫu tuần tự
14.	TDAG	Transition Directed Acyclic Graph	Nén dữ liệu Đồ thị không tuần hoàn có hướng chuyển đổi

15.	W3SVC	World Wide Web Publishing Service.	Một thành phần của Internet Information Services cho phép người dùng xuất bản nội dung lên Internet.

DANH MỤC CÁC KÝ HIỆU TOÁN HỌC

TT	Ký hiệu	Diễn giải
1.	$\langle x, y, z \rangle$	Chuỗi tuần tự có ba phần tử x, y và z
2.	{a, b, c, d}	Tập hợp có 4 phần tử a, b, c và d
3.	sup(X → Y)	Support: Độ hỗ trợ của luật X → Y
4.	conf(X → Y)	Confident: Độ tin cậy của luật X → Y
5.	[l_1, l_2, ..., l_v]	Một dãy có v phần tử l_1, l_2, ..., l_v
6.	O(N^2)	Độ phức tạp N^2
7.	*df*	Chỉ số damping factor: Xác xuất để người dùng tiếp tục truy cập trang Web kế tiếp.
8.	L << N	L rất nhỏ so vớn N
9.	Card{a,b,c,d}	Số lượng phần tử của tập hợp {a, b, c, d}
10.	Count(Arr)	Đếm số phần tử trong mảng Arr
11.	X ⊆ Y	Tập hợp X chứa trong Y (X là tập hợp con của Y)
12.	P(A|B)	Xác xuất của A xuất hiện trong B

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH ẢNH

PHẦN MỞ ĐẦU

1. Giới thiệu

Ngày nay với sự phát triển không ngừng của Công nghệ thông tin và Truyền thông, nó đã ứng dụng vào tất cả các lĩnh vực, đặc biệt là các ứng dụng khai phá dữ liệu trên các Website, trong đó khai phá dữ liệu có tính tuần tự nhằm mục đích dự đoán hành vi truy cập Web là một chủ đề phổ biến, đang được nhiều nhà nghiên cứu quan tâm và mang nhiều ý nghĩa thiết thực. Dự đoán hay phân tích hành vi truy cập web là hướng nghiên cứu gần đây, đóng góp nhiều vào phân tích kinh doanh để phát hiện những dấu hiệu tiềm tàng mới trong hành vi cũng như nhu cầu của khách hàng thương mại điện tử, trò chơi trực tuyến, các ứng dụng web, ứng dụng trên điện thoại di động và IoT. Với lý do đó, nghiên cứu sinh đã quyết định chọn đề tài "Khai phá dữ liệu tuần tự cho dự đoán truy cập Web".

Dự đoán dữ liệu tuần tự là một trong những ứng dụng quan trọng của học máy. Nó được ứng dụng vào việc xây dựng hệ thống khuyến nghị, xử lý ngôn ngữ tự nhiên. Tiềm năng của nó trong khai phá dữ liệu để hỗ trợ ra quyết định là hết sức có ý nghĩa. Những ứng dụng quan trọng và có ý nghĩa ngày nay cho dự đoán chuỗi tuần tự bao gồm dự đoán hành vi truy cập của người dùng; dự đoán ký tự hay từ được gõ trên điện thoại di động, hoặc trên máy tính; dự đoán hành vi mua hàng trên cửa hàng trực tuyến, dự đoán protein kế tiếp trong ngành Sinh Tin học; dự đoán các triệu chứng của bệnh nhân trong bệnh viện, dự đoán thị trường chứng khoán...

Những thử thách đặt ra cho dự đoán chuỗi tuần tự chính là chuẩn hóa dữ liệu để nâng cao hiệu năng và độ chính xác cho việc dự đoán. Hơn nữa việc dự đoán thường được thực hiện trên một không gian dữ liệu khá lớn và cải thiện độ chính xác và thời gian xử lý cho việc dự đoán cũng là các vấn đề rất đáng quan tâm.

Kết quả mong muốn đạt được của nghiên cứu này là một báo cáo đáp ứng yêu cầu cơ bản của luận án tiến sĩ và xuất bản các công trình nghiên cứu (bài báo cũng như hội thảo) liên quan đến nội dung luận án mà được công bố trên các tạp chí uy tín

trong nước và quốc tế. Yêu cầu cần đạt được của luận án là đưa ra tiếp cận hiệu quả hơn các tiếp cận đã có để giải quyết bài toán dự đoán truy cập Web. Cụ thể, trong phạm vi luận án này, các phương pháp mới sẽ được đề xuất như chuẩn hóa cơ sở dữ liệu để phục vụ cho dự đoán, giải pháp nâng cao độ chính xác dự đoán và giải pháp nâng cao hiệu quả về thời gian cho dự đoán.

2. Tính cấp thiết của luận án

Với sự phát triển mạnh mẽ của Internet, nhu cầu người dùng sử dụng Web ngày càng tăng lên để truy cập các thông tin phục vụ cho rất nhiều mục đích khác nhau như tìm tòi, nghiên cứu phục vụ cho học tập, mua sắm, giải trí... theo ước tính của tập đoàn Internet Live Stats (http://www.internetlivestats.com). Các trang Web đã và đang được sử dụng hàng ngày bởi hàng tỷ người. Hơn nữa, World Wide Web là một tài nguyên khổng lồ, đến từ nội dung Web được biểu diễn bởi hàng tỷ trang Web có sẵn trong cộng đồng Internet.

Bên cạnh đó, môi trường Web trong thời đại ngày nay trở thành một môi trường phổ biến cho giao tiếp, tương tác và chia sẻ dữ liệu giữa các người dùng. Điều này dẫn đến hàng ngày, hàng giờ dữ liệu đã không ngừng được tạo ra. Những dữ liệu này có thể được tận dụng để thiết kế và xây dựng các mô hình dự đoán, đặc biệt là mô hình dự đoán hành vi truy cập Web để hỗ trợ ra quyết định. Vấn đề này thực sự rất quan trọng và có ý nghĩa vì dự đoán truy cập Web mang lại nhiều lợi ích cho người sở hữu trang Web cũng như người truy cập Web. Chẳng hạn, đối với người sở hữu trang Web, dự đoán truy cập Web giúp cho họ dự đoán được xu hướng quan tâm của người dùng. Một ví dụ tương tự khác, với một công ty thương mại điện tử trên Internet, dự đoán xu hướng chọn lựa sản phẩm của khách hàng có ý nghĩa rất quan trọng trong chiến lược phát triển sản phẩm của công ty.

Tuy nhiên, sự phát triển không ngừng của các doanh nghiệp hiện đại đã tạo ra áp lực và thách thức không nhỏ cho các nhà nghiên cứu khai phá dữ liệu. Luận án này cố gắng giải quyết những khó khăn này bằng cách đề xuất các mô hình và

giải pháp khai phá dữ liệu tuần tự để dự đoán hành vi truy cập Web hiệu quả hơn như nâng cao độ chính xác và giảm thời gian thực thi dự đoán.

3. Mục tiêu của luận án

Để giải quyết bài toán khai phá dữ liệu tuần tự cho dự đoán truy cập Web, nghiên cứu sinh đề ra 4 mục tiêu chính như sau:

+ Mục tiêu thứ nhất: Nghiên cứu các bài báo liên quan đến luận án để tìm ra những ưu điểm, hạn chế của các bài báo này, từ cơ sở đó nghiên cứu sinh đề xuất các giải pháp tốt hơn cho dự đoán hành vi truy cập Web.

+ Mục tiêu thứ hai: Tìm một mô hình cơ sở dữ liệu phù hợp để hỗ trợ cho dự đoán hành vi truy cập Web.

+ Mục tiêu thứ ba: Tìm giải pháp tốt hơn để nâng cao tính chính xác cho dự đoán hành vi truy cập Web.

Mục tiêu thứ tư: Tìm giải pháp tốt hơn để giảm thời gian thực thi dự đoán hành vi truy cập Web.

Luận án này tập trung vào việc đề xuất mô hình dự đoán khai phá dữ liệu cho dự đoán truy cập Web để nâng cao hiệu quả về độ chính xác và thời gian xử lý cho khai phá dữ liệu mang tính chất tuần tự (còn gọi là dữ liệu phụ thuộc thời gian). Cụ thể là (1) Đề xuất mô hình dự đoán hành vi truy cập web bằng cách tích hợp giải pháp nâng cao độ chính xác và giảm thời gian dự đoán; để triển khai mô hình trên, luận án đưa ra 3 đề xuất tiếp theo là (2) Xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web; (3) Nâng cao hiệu quả thời gian khai phá dữ liệu truy cập tuần tự cho dự đoán truy cập Web; (4) Nâng cao độ chính xác khai phá dữ liệu cho dự đoán truy cập Web.

4. Đối tượng và phạm vi nghiên cứu

Các đối tượng nghiên cứu của luận án bao gồm các lý thuyết về dự đoán dữ liệu tuần tự cho dự đoán truy cập Web như chuỗi tuần tự, cơ sở dữ liệu tuần tự, các tiếp cận dự đoán truy cập Web, các bộ dữ liệu click-stream phục vụ cho khai phá dữ

liệu tuần tự như MSNBC, FIFA, KOSARAK [1], các bộ dữ liệu Weblog (palmviewsanibel.com [2] , periwinklecottages.com [3], inees.org [4], lvtm.vn [5]…) các giải thuật hỗ trợ cho dự đoán (CPT+, PageRank, phân tích và xử lý chuỗi…), các bài báo liên quan đến đề tài được xuất bản trên các tạp chí, các hội nghị khoa học trong nước hoặc quốc tế.

Phạm vi nghiên cứu của luận án là khai phá dữ liệu tuần tự cho dự đoán truy cập Web trên các tập clickstream và dữ liệu nhật ký truy cập Web (Web Log) lưu trên các máy chủ Web, cụ thể là dữ liệu nhật ký thuộc các Web Server như IIS (máy chủ Web trên hệ điều hành Microsoft Windows) và Apache (Các máy chủ Web trên các Hệ điều hành họ Linux).

5. Các vấn đề nghiên cứu

Vấn đề nghiên cứu 1: Để khai phá dữ liệu tuần tự cho dự đoán truy cập Web cần có những mô hình nào?

Nội dung chi tiết của Chương 1 trong luận án sẽ trả lời cho ***vấn đề nghiên cứu 1***

Vấn đề nghiên cứu 2: Cơ sở dữ liệu tuần tự cho dự đoán hành vi truy cập Web được xây dựng như thế nào?

Nội dung chi tiết của Chương 2 trong luận án sẽ trả lời cho ***vấn đề nghiên cứu 2***

Vấn đề nghiên cứu 3: Làm thế nào để nâng cao độ chính xác cho dự đoán truy cập Web dùng mô hình dự đoán chuỗi tuần tự theo mô hình cây dự đoán nén (Compact Prediction Tree - CPT+) ?

Nội dung chi tiết của Chương 3 trong luận án sẽ trả lời cho ***vấn đề nghiên cứu 3***

---

[1] *http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php*

[2] *Truy cập ngày 29/08/2017*

[3] *Truy cập ngày 22/08/2017*

[4] *Truy cập ngày 25/08/2017*

[5] *Truy cập ngày 12/06/2018*

Vấn đề nghiên cứu 4: Làm thế nào để giảm thời gian thực thi dự đoán cho dự đoán truy cập Web dùng mô hình dự đoán chuỗi tuần tự theo mô hình cây dự đoán nén (Compact Prediction Tree - CPT+) ?

Nội dung chi tiết của Chương 4 trong luận án sẽ trả lời cho ***vấn đề nghiên cứu 4***

6. Phương pháp nghiên cứu

Về hướng tiếp cận, luận án theo hướng tiệp cận Cây dự đoán nén CPT+ (Compact Prediction Tree) để xây dựng các mô hình và giải pháp dự đoán hành vi truy cập web tăng độ chính xác hoặc/và giảm thời gian xử lý. Để mô hình đề xuất dựa trên CPT+ nâng cao hiệu năng cho dự đoán hành vi truy cập Web, nghiên cứu sinh đã nghiên cứu tích hợp thêm giải thuật PageRank và kỹ thuật xử lý chuỗi.

Trong quá trình nghiên cứu luận án, nghiên cứu sinh đã sử dụng những phương pháp nghiên cứu như sau:

+ Phương pháp thu thập dữ liệu: Các bộ dữ liệu click-stream, các bộ dữ liệu Weblog và dữ liệu đặt hàng trong thương mại điện tử được sử dụng trong luận án là dữ liệu thứ cấp và được thu thập từ các nguồn dữ liệu khác nhau có nguồn gốc rõ ràng, khách quan và ghi nhận nhật ký truy cập Web.

+ Phương pháp hỏi ý kiến chuyên gia: Trước và trong thời gian thực hiện luận án, nghiên cứu sinh đã liên hệ với nhiều chuyên gia, các nhà nghiên cứu để được góp ý về tên đề tài cũng như nội dung cần nghiên cứu. Trong đó, vai trò định hướng và góp ý của GS.TS Philippe Fourier Viger [1] là rất quan trọng.

+ Phương pháp nghiên cứu định lượng: Nghiên cứu sinh tiến hành nghiên cứu thử nghiệm có hệ thống về các hiện tượng quan sát được qua các số liệu thống kê, toán học và thông qua việc phát triển các giải thuật như các giải thuật về xây dựng cơ sở dữ liệu tuần tự, tính toán PageRank, kỹ thuật xử lý chuỗi...

---

[1] Chuyên gia về Data Mining, Big Data, Artificial Intelligence, Pattern Mining, Itemset Mining, Graph Mining, Sequence Prediction, công tác tại Harbin Institute of Technology, China (http://www.philippe-fournier-viger.com/publications.php).

+ Phương pháp nghiên cứu định tính: Nghiên cứu sinh tiến hành đánh giá các giải pháp đề xuất như so sánh các phương pháp mới và cải tiến cho dự đoán truy cập Web về phương diện thời gian và độ chính xác để xem xét giải pháp đề xuất có phù hợp hay không, chẳng hạn như có ý nghĩa về mặt thống kê hay không.

+ Phương pháp nghiên cứu phân tích và tổng hợp: Nghiên cứu, tìm hiểu và tổng hợp các lý thuyết liên quan đến đề tài như lý thuyết về dự đoán tuần tự, thuật toán CPT (Compact Prediction tree), thuật toán PageRank. Bên cạnh việc nghiên cứu lý thuyết, nghiên cứu sinh cũng tìm hiểu các nghiên cứu liên quan đến luận án để phân tích điểm yếu, điểm mạnh của các phương pháp dự đoán truy cập Web. Từ việc phân tích và tổng hợp đó, nghiên cứu sinh có cơ sở để đề xuất các giải pháp tốt hơn cho dự đoán truy cập Web so với các tiếp cận thông thường.

7. Các đóng góp của luận án

Các đóng góp cho dự đoán truy cập Web được trình bày trong luận án và các công trình nghiên cứu liên quan của nghiên cứu sinh bao gồm các nội dung chính sau:

- 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-Fold Check 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 40% (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 60% 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 0.024% và thời gian dự đoán nhanh hơn xấp xỉ 60 lần so với tiếp cận thông thường.

8. Bố cục của luận án

Bố cục luận án gồm có năm chương và một phần kết luận. Cụ thể, trong chương đầu tiên, nghiên cứu sinh trình bày tổng quan về vấn đề cần nghiên cứu. Ở chương tiếp theo, nghiên cứu sinh đưa ra các khái niệm về dữ liệu tuần tự và trình bày phương pháp thiết kế cơ sở dữ liệu tuần tự để dự đoán truy cập Web. Trong Chương 3, nghiên cứu sinh trình bày về giải pháp nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Tiếp theo, trong Chương 4, nghiên cứu sinh đề xuất giải pháp nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Bên cạnh đó, trong Chương 5, nghiên cứu sinh trình bày giải pháp tích hợp nâng cao độ chính xác và nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Danh mục các chương của luận án như sau:

Chương 1. Tổng quan về khai phá dữ liệu tuần tự cho dự đoán truy cập Web.

Chương 2. Chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web.

Chương 3. Nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web.

Chương 4. Nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web.

Chương 5. Tích hợp nâng cao độ chính xác và nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web.

Phần cuối cùng là phần kết luận, trong phần này nghiên cứu sinh sẽ rút ra các kết luận và trình bày hướng phát triển của luận án.

CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU TUẦN TỰ CHO DỰ ĐOÁN TRUY CẬP WEB

1.1. Giới thiệu

Ngày nay với sự phát triển không ngừng của Công nghệ thông tin và Truyền thông, vấn đề dự đoán và hỗ trợ ra quyết định trong khai phá dữ liệu là một chủ đề rất quan trọng và hữu ích. Trong đó, dự đoán hành vi truy cập Web với nhiều nghiên cứu đã được thực hiện gần đây có nhiều ảnh hưởng tích cực trên nhiều lãnh vực khác nhau của ứng dụng Web như các hệ thống khuyến nghị (Recommender System), các hệ thống dự đoán (Prediction System), các hệ thống bảo mật mạng (Cyber Security System). Các nghiên cứu này được thực hiện hầu hết dựa vào các giải thuật sử dụng dữ liệu lịch sử truy cập Web, còn gọi là các Web log, ghi nhận thông tin liên quan đến truy cập của người dùng Web.

Nhiều nghiên cứu dự đoán truy cập Web với tiếp cận dựa trên máy học để đề xuất các phương pháp khác nhau. Chẳng hạn như các nghiên cứu dùng phương pháp Machine Learning [10; 15; 52; 107], Association Rules [43; 60; 71; 74; 113], Sequential Pattern [79], Sequential Rules [36; 40; 44], Markov [25; 26; 57; 82; 83; 95; 110]. Bên cạnh đó, nhiều phương pháp lai cũng được áp dụng như các nghiên cứu kết hợp Markov với Clustering [8; 9; 100; 109; 119], phương pháp kết hợp Markov, Clustering và Association Rules [27; 62; 119], hay phương pháp kết hợp Markov, Clustering và Sequential Pattern Mining [7]. Nhìn chung, các phương pháp này đã sử dụng phương pháp khai phá mẫu và tìm ra các luật hợp lệ rồi đưa ra kết quả dự đoán với độ tin cậy và độ hỗ trợ cho các luật. Tuy nhiên, những phương pháp này có những hạn chế như sự bùng nổ các luật do đó tốn bộ nhớ và thời gian xử lý rất chậm. Hơn nữa, độ chính xác cũng không đảm bảo vì tiếp cận này đã dựa trên mô hình làm mất thông tin và bỏ qua các trường hợp hiếm. Độ chính xác được xác định bằng công thức [48]:

$$Accuracy = |successes| / |sequences| \quad (1.1)$$

Trong đó

Accuracy: Độ chính xác của dự đoán

|*successes*|: Số lượng chuỗi dự đoán thành công

|*sequences*|: Số lượng chuỗi dự đoán

Các khai phá dữ liệu dựa vào luật để dự đoán theo giải thuật CMRules và CMDEO [32] sử dụng tiếp cận « generate-candidate-and-test » mà có thể phát sinh một khối lượng lớn các ứng viên, một mặt điều này làm tốn thời gian xử lý và không gian nhớ, mặt khác nhiều ứng viên đã bị bỏ qua do đó dẫn đến sự mất thông tin trong quá trình khai phá luật. Bên cạnh đó, sự bùng nổ tổ hợp của các tiếp cận dựa theo Markov, mà điển hình là mô hình All-K-Markov cũng làm chậm quá trình dự đoán. Mặc dù có rất nhiều nghiên cứu để cải thiện hạn chế này về mặt NÂNG CAO HIỆU QUẢ THỜI GIAN xử lý nhưng độ chính xác lại giảm đi [46].

Các ứng dụng điển hình cho dự đoán chuỗi dữ liệu tuần tự đã được các nhà nghiên cứu đề xuất trong thời gian rất gần đây như dự đoán hành trình xe lửa [55], dự đoán hành vi người dùng trong truyền thông [30], khai phá dữ liệu tuần tự trong lãnh vực y tế [98; 99], phân tích dữ liệu liên tục ứng dụng cho xây dựng mô hình thông tin [92], dự đoán hành vi người học [1] …

Do vậy, dự đoán hành vi truy cập Web của người dùng dựa trên dự đoán chuỗi dữ liệu tuần tự là một tiếp cận hứa hẹn và mở ra nhiều cơ hội nghiên cứu. Phần tiếp theo trình bày các khái niệm về chuỗi tuần tự truy cập Web, cơ sở dữ liệu tuần tự truy cập Web và dự đoán hành vi truy cập Web. Dựa trên các kết quả đã được tìm hiểu, nghiên cứu và khảo sát, chương này sẽ giới thiệu vấn đề khai phá dữ liệu tuần tự để dự đoán hành vi truy cập Web một cách hình thức trong trường hợp tổng quát. Bên cạnh đó, các ưu điểm, hạn chế của các phương pháp truyền thống liên quan đến vấn đề này cũng sẽ được trình bày. Hơn nữa, một phương pháp

mới, CPT+, hiệu quả hơn về độ chính xác và độ phức tạp về thời gian so các phương pháp truyền thống.

1.2. Khái niệm dự đoán hành vi truy cập Web

Định nghĩa 1.1

Gọi $U = \{IP_1, IP_2, ..., IP_k\}$ là tập hợp người dùng truy cập Web với IP_i là địa chỉ IP của người dùng truy cập thứ i ($1 \leq i \leq k$) và k là số lượng của các địa chỉ IP.

Cho một tập hợp các phần tử hữu hạn (ký hiệu) $I = \{i_1, i_2, ..., i_m\}$, một chuỗi tuần tự Seq là một danh sách có thứ tự $Seq = \langle p_1, p_2, ... p_n \rangle$, trong đó $p_x \in I$ ($1 \leq x \leq n$).

Gọi $S = \langle p_1, p_2, ... p_q \rangle$, $S \in Seq$ là chuỗi tuần tự các trang Web được truy cập bởi người dùng có địa chỉ IP_i với $IP_i \in U$ và q là số lượng của các trang Web được truy cập.

Nhật ký truy cập Web $L = [l_1, l_2, ..., l_v]$ là một dãy các dòng nhật ký l_j ($1 \leq j \leq v$) với v là số dòng nhật ký và $l_j = (IP_i, p_i, t_i)$ là dòng nhật ký thứ j ghi nhận người dùng có địa chỉ $IP_i \in U$, truy cập vào trang Web $p_i \in S$ vào thời điểm t_i.

Trên đây là định nghĩa nhật ký truy cập Web mang tính hình thức để phục vụ cho các nghiên cứu liên quan của luận án. Nhật ký máy chủ Web Log đầy đủ là một danh sách các thành phần nhật ký mà tuân theo chuẩn CLF (Common Log Format) gồm các thuộc tính: địa chỉ IP máy khách hay hostname, thời gian truy cập, phương thức HTTP yêu cầu, đường dẫn của nguồn tài nguyên được truy cập trên máy chủ Web (nhận dạng URL), nghi thức được dùng (HTTP/1.0, HTTP/1.1), mã số tình trạng, số lượng byte được truyền, referer, user-agent, ... Thuộc tính referer trình bày URL mà người dùng đã truy cập đến trang được yêu cầu. Thuộc tính user-agent là phần mềm được dùng để truy cập các trang, có thể là Web Clawler (ví dụ GoogleBot, openbot, scooter, ...) hay một trình duyệt (Google Chrome, Mozilla, Internet Explorer, Opera, ...).

Định nghĩa 1.2

Cơ sở dữ liệu tuần tự truy cập Web $SD = \{s_1, s_2, \ldots, s_N\}$ là tập hợp các chuỗi $s_m \in S$ ($1 \leq m \leq N$) với N là số lượng các chuỗi dữ liệu tuần tự trong cơ sở dữ liệu tuần tự này.

Chẳng hạn, ***Bảng 1.1*** trình bày một cơ sở dữ liệu tuần tự truy cập Web chứa 5 chuỗi tuần tự được truy cập bởi 5 người dùng có địa chỉ IP khác nhau. Trong đó, chuỗi tuần tự truy cập Web thứ nhất có 6 trang Web p_1, p_2, p_4, p_6, p_3 và p_5 được truy cập bởi người dùng có địa chỉ IP_1 theo thứ tự thời gian. Tương tự, chuỗi tuần tự truy cập Web thứ hai thể hiện người dùng có địa chỉ IP_2 truy cập lần lượt vào các trang Web p_4, p_3, p_5, p_6, p_2.

Bảng 1.1 *Một ví dụ về cơ sở dữ liệu tuần tự truy cập Web*

Địa chỉ IP	**Chuỗi tuần tự truy cập Web**	
IP_1	s_1	$\langle p_1, p_2, p_4, p_6, p_3, p_5 \rangle$
IP_2	s_2	$\langle p_4, p_3, p_5, p_6, p_2 \rangle$
IP_3	s_3	$\langle p_1, p_2, p_4, p_9, p_3, p_7, p_{10} \rangle$
IP_4	s_4	$\langle p_6, p_1, p_4, p_8, p_3, p_5 \rangle$
IP_5	s_5	$\langle p_4, p_2, p_8, p_6, p_3, p_5 \rangle$

Định nghĩa 1.3

Cho một chuỗi tuần tự các trang Web cần được dự đoán trang Web truy cập kế tiếp $S_{query} = \langle page_1, page_2, \ldots page_m \rangle$, $S_{query} \in Seq$ và $page_i$ là trang Web được truy cập thứ i ($1 \leq i \leq m$) và m là số lượng các trang Web trong chuỗi S_{query} (m còn được gọi là chiều dài của chuỗi S_{query}).

Dự đoán hành vi truy cập Web là dự đoán trang Web sẽ được truy cập kế tiếp p_{next} của S_{query} trên cơ sở dữ liệu tuần tự truy cập Web SD bằng cách sử dụng phương pháp dự đoán chuỗi tuần tự truy cập Web, chẳng hạn như phương pháp dự

đoán chuỗi dữ liệu tuần tự [46] và việc dự đoán hành vi truy cập Web này được đặc tả bằng công thức sau:

$$P_{next} = F\ (S_{query}, SD) \qquad (1.2)$$

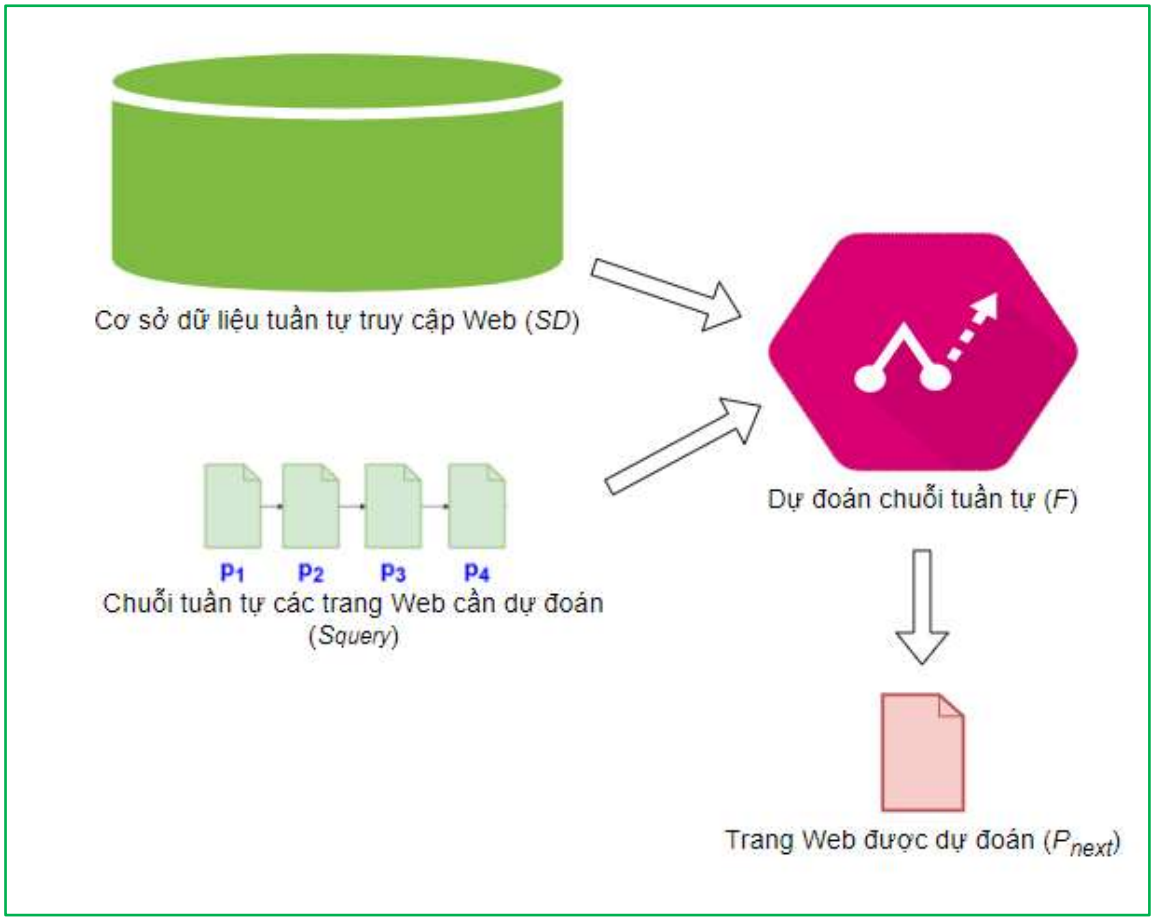
Trong đó:

P_{next} là trang Web kế tiếp được dự đoán.

F hàm xử lý dự đoán.

S_{query} là chuỗi tuần tự các trang Web cần dự đoán.

SD là cơ sở dữ liệu tuần tự truy cập Web. ***Hình 1.1*** trình bày mô hình phổ biến cho dự đoán truy cập Web.



Hình 1.1 *Mô hình phổ biến cho dự đoán truy cập Web*

Trong một số nghiên cứu trước đây *F* có thể dùng độc lập hay kết hợp nhiều nhiều pháp như: Luật kết hợp, Clustering, Compact Prediction Tree (CPT), Compact Prediction Tree Plus (CPT+). Những ưu điểm và hạn chế của các mô hình này sẽ được trình bày ở phần 1.4.3.

1.3. Các phương pháp phổ biến

Theo F. Khalil và các đồng sự [63], những phương pháp phổ biến để dự đoán truy cập Web là khai phá bằng luật kết hợp (Association Rules), gom cụm (Clustering) và mô hình xác suất Markov. Bên cạnh đó, [46] đã chỉ ra những hạn chế của các phương pháp trên và trình bày một phương pháp mới để khai phá chuỗi dữ liệu tuần tự, trong đó một trong những ứng dụng quan trọng là dự đoán hành vi truy cập Web.

1.3.1. Phương pháp luật kết hợp

1.3.1.1. Khái niệm

Định nghĩa 1.5: *Khai phá luật kết hợp* [38]

Khai phá luật kết hợp *Association rule mining (ARM)* là phương pháp tìm các luật kết hợp trong một cơ sở dữ liệu các giao tác. Một luật kết hợp là một luật có dạng $X \rightarrow Y$, trong đó, X và Y là hai tập hợp các phần tử sao cho $X \cap Y = \emptyset$. Các luật kết hợp được đánh giá bằng cách dùng các độ đo sự quan tâm. Các độ đo thường được dùng trong ARM gồm có:

- ✓ Độ hỗ trợ : $sup(X \rightarrow Y) = sup(X \cup Y)$

  Trong đó
- ✓ Độ tin cậy: $conf(X \rightarrow Y) = sup(X \cup Y)/sup(X)$
- ✓ Chỉ số đo lường: $lift(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X) \times sup(Y)}$

Theo các nghiên cứu [39; 75], *độ hỗ trợ* được xác định bằng tỷ số giữa số lượng các chuỗi mà X xuất hiện phía trước Y chia cho tổng số các chuỗi và được xác định bằng công thức $sup(X \rightarrow Y) = P(X \cup Y)$ hoặc $sup(X \cup Y) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$ với

T là tập các giao tác của một cơ sở dữ liệu cho trước [56]; *độ tin cậy* được xác định bằng tỷ số giữa số lượng các chuỗi mà X xuất hiện trước Y chia cho số lượng chuỗi mà X xuất hiện, được xác định bằng công thức $P(Y|X) = P(X \cup Y)/P(X)$. Để xác định các luật kết hợp, người dùng phải cung cấp một ngưỡng hộ trợ tối thiểu và một ngưỡng tin cậy tối thiểu.

Bên cạnh đó, *lift* là chỉ số đo lường được dùng để lọc các luật thừa khi các vế trái và vế phải của các luật được phân tích và được tính bằng công thức $P(X \cup Y)/ P(X) P(Y)$. Nếu chỉ số $lift > 1$ cho chúng ta biết về mức độ mà hai sự kiện phụ thuộc vào một sự kiện khác, điều này hữu ích cho dự đoán . Nếu chỉ số $lift < 1$ cho chúng ta biết các phần tử thay thế nhau, điều này ảnh hưởng tiêu cực đến sự hiện diện của phần tử khác và ngược lại.

1.3.1.2. Các công trình nghiên cứu liên quan

Công trình [43] dự đoán hành vi truy cập của người dùng vào các trang Web trên tập dữ liệu truy cập Web MSNBC. Đây là tập dữ liệu khá lớn và số lượng luật thu được cũng khá nhiều. Trong nghiên cứu này giải thuật Priori-PT đã được áp dụng để khai phá các luật. Các luật dư thừa đã được loại bỏ dựa trên ngưỡng hỗ trợ tối thiểu. Các kết quả dự đoán cho thấy rằng sự quan tâm của người dùng đến các trang phụ thuộc chủ yếu độ hỗ trợ và chỉ số lift do đó điều này không những tốn chi phí thời gian khi xử lý các mẫu số có số lượng lớn mà còn không sử dụng thông tin sẵn có trong các chuỗi dữ liệu để thực hiện dự đoán một cách chính xác.

Để cải tiến khai phá luật kết hợp, một số nghiên cứu cải tiến luật kết hợp cũng đã được đề xuất như sau:

- Nghiên cứu cải tiến luật kết hợp CMRules [32]: Giải thuât CMRules thực thi bằng cách tìm các luật kết hợp để cắt tỉa không gian tìm kiếm cho các phần tử xuất hiện cùng với nhau trong nhiều chuỗi tuần tự. Sau đó, nó loại bỏ các luật kết hợp mà nhỏ hơn các ngưỡng độ tin cậy tối thiểu và độ hỗ trợ tối thiểu theo thứ tự thời gian.

- Theo [39], một cải tiến của luật kết hợp là khai phá luật tuần tự, theo đó, một luật tuần tự $X \rightarrow Y$ là một mối quan hệ giữa hai tập hợp các phần tử không rỗng và không giao nhau X, Y. Luật dạng này có hai đặc tính: Độ hỗ trợ: *[Số các chuỗi tuần tự mà X xuất hiện trước Y] / [Tổng số các chuỗi tuần tự]*; Độ tin cậy: *[Số các chuỗi tuần tự mà X xuất hiện trước Y] / [Tổng số các chuỗi tuần tự mà X xuất hiện]*. Nhiệm vụ của khai phá luật liên tiếp là tìm tất cả các luật hợp lệ với một độ hỗ trợ và độ tin cậy không thấp hơn các ngưỡng *minSup* và *minConf* mà người dùng định nghĩa trước. Các nghiên cứu điển hình của khai phá luật tuận tự là [32; 37; 39].
- Nghiên cứu cải tiến luật kết hợp RuleGrowth [39]: Hầu như các nghiên cứu trước đó về luật kết hợp tập trung vào việc tìm các luật xuất hiện trong một chuỗi đơn các sự kiện. Giải thuật RuleGrowth giải quyết tốt hạn chế này. Sử dụng tiếp cận tăng trưởng mẫu đã đem lại hiệu quả và độ linh hoạt cho giải thuật này.
- Nghiên cứu cải tiến luật kết hợp ER-Miner: ER-Miner dựa trên ý tưởng tìm các lớp tương đương của các luật có cùng vế trái và vế phải và có ưu điểm là giải quyết các vấn đề của các nghiên cứu về luật tuần tự trước đó như phép chiếu cơ sở dữ liệu tốn chi phí về thời gian và hiệu năng thấp khi khai phá cơ sở dữ liệu chứa các chuỗi dữ liệu dài [37] . Bên cạnh đó, ER-Miner cũng dùng một cấu trúc gọi là SCM (Spare Count Matrix) để làm gọn không gian tìm kiếm.

1.3.1.3. Ưu điểm và hạn chế

- ***Ưu điểm:*** Nghiên cứu khai phá luật kết hợp là một trong những nghiên cứu khai phá dữ liệu hiệu quả. Đây là tiếp cận tìm các mối quan hệ giữa các thuộc tính trong cơ sở dữ liệu và tạo ra các phát biểu có dạng if-then liên quan đến các giá trị thuộc tính [41] . Hơn nữa, các luật kết hợp có thể được

dùng để quyết định các truy cập kế tiếp dựa trên các mối tương quan thống kê có ý nghĩa [21].

- ***Hạn chế:*** Rất tốn chi phí thời gian khi xử lý các mẫu số có số lượng lớn và dài và được xây dựng trên mô hình không hỗ trợ dự đoán nên không sử dụng thông tin sẵn có trong các chuỗi dữ liệu để thực hiện dự đoán. Bên cạnh đó thứ tự về thời gian của các trang Web được dự đoán bị bỏ qua trong quá trình xử lý [38]. Ngoài ra, phương pháp này không sử dụng thông tin chứa trong các chuỗi dữ liệu tuần tự do đó làm giảm độ chính xác [38].

1.3.2. Phương pháp chuỗi Markov

1.3.2.1. Khái niệm

Nghiên cứu [26] cho rằng sử dụng các mô hình chuỗi Markov truyền thống, các nhà nghiên cứu so sánh các phần tử của các tiền tố tối đa của chuỗi tuần tự các truy cập Web dựa theo lịch sử truy cập với các phần tử từ các hậu tố có cùng chiều dài của chuỗi tuần tự truy cập hiện tại của người dùng và nhận được các chuỗi tuần tự theo lịch sử truy cập với xác xuất cao nhất của các phần tử về mặt so khớp các phần tử. Ngoài ra, theo nghiên cứu [26], mô hình 0-order Markov là mô hình xác suất dựa vào tỷ lệ không có điều kiện:

$$p(x_n) = Pr(X_n) \qquad (1.3)$$

Trong đó, $p(x_n)$ là xác suất của một lần viếng thăm trang. Điều này có thể được ước tính như là tỷ lệ của số lần viếng thăm đến một trang trong một khoảng thời gian nào đó [93]. Bên cạnh đó, công trình [93] trình bày mô hình 1-order Markov được sử dụng bởi [12] và [84] đã đề cập đến xác suất chuyển đổi từng trang này sang trang khác. Những điều này có thể được ước tính từ các n-gram của bộ $<X_1, X_2>$ để đưa ra xác xuất có điều kiện:

$$p(x_2 \mid x_1) = Pr(X_2 = x_2 \mid X_1 = x_1) \qquad (1.4)$$

Để ghi nhận các đường dẫn dài hơn của các truy cập, [93] xét đến xác xuất có điều kiện mà các chuyển đổi trang của một người dùng truy cập đến một trang thứ n có số lần truy cập trước đó là k = n -1:

$$p(x_n \mid x_{n-1},...x_{n-k}) = Pr(X_n = x_n \mid X_{n-1},...,X_{n-k}) \quad (1.5)$$

Những xác xuất có điều kiện như vậy được xem là các xấp xỉ K^{th}-order Markov (hay còn gọi là các mô hình K^{th}-order Markov).

Theo nghiên cứu [63], phương pháp Markov cho dự đoán Web được phân thành 3 loại chính như sau:

(1) Mô hình All-K^{th} Order Markov: Đối với mỗi tiến trình thử nghiệm, mô hình chuỗi Markov thứ tự cao nhất kể cả tiến trình được dùng để dự đoán tiến trình.

(2) Mô hình Frequency pruned Markov:

Các trạng thái có tần số thấp được loại bỏ. Sự loại bỏ các trạng thái thấp này ảnh hưởng đến độ chính xác của mô hình chuỗi Markov. Tuy nhiên, số lượng các trạng thái của mô hình được cắt tỉa Markov sẽ được giảm đáng kể.

(3) Mô hình Accuracy prunced Markov: Các trạng thái với độ độ chính xác dự đoán thấp có thể bị loại bỏ. Một cách để ước tính độ chính xác dự đoán dùng xác suất điều kiện được gọi là quá trình cắt tỉa tin cậy. Một cách khác để ước tính độ chính xác dự đoán là đếm (ước lượng) các lỗi liên quan, gọi là quá trình cắt tỉa lỗi. Việc đánh giá của quá trình cắt tỉa đã cho thấy rằng có đến 90% các trạng thái có thể được cắt tỉa dẫn đến độ phức tạp không gian trạng thái giảm xuống và làm tăng độ bao phủ mà độ chính xác vẫn không thay đổi. Độ phức tạp không gian trạng thái là một vấn đề chính khi thực hiện mô hình chuỗi Markov. Các thứ tự cao hơn dẫn đến nhiều trạng thái hơn nhưng chúng thường đưa ra độ chính xác dự đoán tốt hơn vì chúng quan sát lịch sử truy cập trước đó. Một khó khăn khác mà nảy sinh khi xây dựng các mô hình chuỗi Markov là cần thiết để thu được độ chính xác dự đoán tốt hơn, chúng được kết hợp với độ phức tạp không gian trạng thái cao hơn.

1.3.2.2. Các nghiên cứu liên quan

Theo nghiên cứu [26], các mô hình chuỗi Markov truyền thống dự đoán truy cập Web bằng cách so khớp chuỗi truy cập tuần tự hiện tại của người dùng với các chuỗi truy cập của các chuỗi tuần tự truy cập trong lịch sử truy cập trước đó. Các nghiên cứu điển hình sử dụng mô hình truyền thống này là các nghiên cứu [84; 91; 101].

Nghiên cứu [102] đã đề xuất giải thuật cho định hướng đường dẫn (tour generation - TUMMs) dựa vào mô hình chuỗi Markov. Ma trận chuyển đổi trạng thái của mô hình chuỗi Markov có thể được xem như một biểu diễn "chuyển đổi trọng số". Họ đã giới thiệu một ứng dụng về thủ tục ước lượng trọng số Hub/Authority để tạo ra thông tin truy cập người dùng cá nhân hóa.

Bài báo [26] trình bày mô hình HTMM để dự đoán truy cập Web. Đây là một mô hình cấu trúc giống cây mà kết hợp phương pháp dự đoán các chuỗi tuần tự truy cập bằng cách so khớp và một phương pháp chuỗi Markov lai.

Nghiên cứu [119] sử dụng mô hình Markov để đưa ra dự đoán liên kết hỗ trợ người dùng mới để điều hướng trang web dựa trên hành vi của người dùng truy cập Web trong quá khứ. Các tác giả của công trình này dùng một giải thuật nén ma trận thống kê trạng thái để phân nhóm các trang Web theo hành vi trạng thái tương tự và nén ma trận trạng thái để thu được kích cỡ tối ưu cho tính toán thống kê hiệu quả nhằm dự đoán các liên kết.

Nghiên cứu [25] đã đề xuất hai tiếp cận để dự đoán truy cập Web: Tiếp cận thứ nhất sử dụng mô hình Morse [81], và tiếp cận thứ hai là chuỗi Markov ergodic (các đặc trưng thống kê của nó có thể suy ra được từ một chuỗi các mẫu đủ dài của nó).

Nghiên cứu [21] so sánh và phân tích 3 mô hình 1st-order Markov, 2nd-Markov dùng cho dự đoán truy cập Web trên cơ sở dữ liệu Web Log được truy cập

bởi người dùng. Kết quả nghiên cứu cho rằng 1st-Markov là mô hình dự đoán tốt nhất khi so sánh 3 mô hình này.

Nghiên cứu [95] giới thiệu mô hình gọi là HSMP để dự đoán các truy cập Web theo chủ đề theo hai giai đoạn: (1) Sử dụng yếu tố Relevance mà có thể được sử dụng để suy luận người dùng hành vi duyệt giữa các danh mục web; (2) Dự đoán các chủ đề truy cập Web theo cách kết hợp các mô hình chuỗi Markov khác nhau một cách hợp lý.

Nghiên cứu [9] trình bày mô hình all-K^{th} Markov cải tiến cho dự đoán truy cập Web để làm hạn chế khả năng mở rộng về số lượng đường dẫn mà không ảnh hưởng đến độ chính xác.

1.3.2.3. Ưu điểm và hạn chế

- ***Ưu điểm:***
  - Những mô hình chuỗi Markov là một trong những mô hình khả thi để mô hình hóa chuỗi tuần tự Web. Các mô hình chuỗi Markov có thể được ước lượng theo thống kê, thích nghi và mang tính phát sinh do vậy mô hình chuỗi Markov giúp cho dự đoán truy cập Web và định hướng đường dẫn truy cập [102].
  - Các mô hình chuỗi Markov và các biến thể của chúng hay các mô hình dựa trên chuỗi tuần tự rất thích hợp cho dự đoán truy cập Web [50].
- ***Hạn chế:***

- Mô hình 1st – Markov không xem xét đến yếu tố "long-term memory" của hành vi truy cập Web của người dùng vì dựa trên giả định rằng trạng thái kế tiếp của hành vi truy cập chỉ đơn giản là một hàm của trạng thái hiện tại [31].

- Các mô hình Lower-order Markov không thể thành công để dự đoán truy cập Web vì chúng không đủ tin cậy đoán hành vi truy cập của người dùng trong quá khứ [26]. Bên cạnh đó, các mô hình Lower-order không dùng đủ lịch sử duyệt Web của người dùng và do vậy thiếu tính chính xác [61].

- Các mô hình Higher-order Markov có kết quả dự đoán tốt hơn Lower-oder tuy nhiên độ phức tạp về không gian khá cao và có độ bao phủ (coverage) thấp [26]. Hơn nữa, theo [78], phương pháp Higher-order Markov có những hạn chế như: (1) Độ phức tạp về trạng thái, (2) Độ bao phủ không cao; Thỉnh thoảng cho ra kết quả không chính xác về dự đoán. Hơn nữa theo [31], độ phức tạp của không gian trạng thái Higher-order Markov theo hàm mũ; Thứ tự của mô hình này càng cao thì độ phức tạp càng cao. Điều này làm độ chính xác giảm xuống đáng kể.
- Theo nghiên cứu [46], các mô hình chuỗi Markov có những giới hạn như: (1) Mô hình chuỗi Markov dựa trên cơ sở là mỗi sự kiện đơn lẻ phụ thuộc vào các sự kiện trước do vậy độ chính xác của mô hình này sẽ không cao; (2) Chỉ một phần thông tin của chuỗi tuần tự được xem xét; (3) Việc tăng thứ tự k trong mô hình K^{th}-order Markov sẽ dẫn đến độ phức tạp về thời gian tăng lên.

1.3.3. Phương pháp Clustering

1.3.3.1. Khái niệm

Theo nghiên cứu [63], động cơ chính đằng sau cách sử dụng Clustering là để cải tiến hiệu quả và khả năng mở rộng của các nhiệm vụ cá nhân hóa theo thời gian thực. Nói chung, Clustering nhắm đến việc chia dữ liệu thành nhiều nhóm (hay còn gọi là các Clustering) trong đó các độ tương tự bên trong được giảm tối thiểu trong khi các độ tương tự của mỗi nhóm được tăng tối đa. Các phiên truy cập Web có thể đạt được thông qua phân cụm trang hay phân cụm người dùng. Phân cụm trang được thực hiện bằng cách nhóm các trang có nội dung tương tự. Mặt khác, phân cụm truy cập người dùng liên quan đến việc chọn mô tả dữ liệu thích hợp cho một phiên truy cập người dùng và định nghĩa độ tương tự giữa hai phiên truy cập.

Các thuật toán liên quan đến Clustering gồm có K-Means, DBScan, Bisecting K-Means, Hierarchical Clustering…

1.3.3.2. Các nghiên cứu liên quan, ưu điểm và hạn chế

Nghiên cứu [111] đề xuất hệ khuyến nghị dùng giải thuật K-means để tìm các mẫu truy cập tương tự của các truy cập người dùng. Ngoài ra, để phân lớp và dự đoán, giải thuật KNN (k-nearest neighbor) được thực hiện. Tiếp cận này cũng kết hợp dữ liệu mẫu truy cập người dùng mà thuộc về người dùng khác do đó mô hình được đề xuất cũng dự đoán các mẫu truy cập xuất hiện không thường xuyên. Như vậy, việc đưa ra các khuyến nghị dữ liệu lịch sử Web là cá nhân hóa, phụ thuộc vào những tần số liên kết URL, các tần số truy cập của người dùng, phân tích dữ liệu theo phiên truy cập và phân tích dữ liệu theo thời gian. Hơn nữa để kết hợp các tham số này một kỹ thuật gán trọng số được sử dụng.

Nghiên cứu [49] trình bày một mô hình dự đoán mà gồm ba thành phần khác nhau. Đầu tiên, các tác giả đề xuất một tối ưu theo thời gian chờ được tinh chỉnh để nhận dạng phiên truy cập. Thứ hai, họ đã đề xuất cách sử dụng một giải thuật dựa trên mật độ cụ thể để khám phá mẫu truy cập Web. Các giải pháp dự đoán truy cập trực tuyến dựa trên Clustering không chính xác bằng tiếp cận k-Nearest-Neighbors (kNNs); tuy nhiên kNN có vấn đề về khả năng mở rộng. Trong quá trình khai phá tinh chỉnh được đề xuất của họ, họ đề nghị một phương pháp kNN nhanh hơn để dự đoán trực tuyến các mẫu truy cập. Cuối cùng, một tiếp cận mới cho dự đoán trực tuyến hiệu quả cũng được đề xuất. Các kết quả thử nghiệm mô tả khả năng ứng dụng và độ hiệu quả của tiếp cận được đề xuất.

Nghiên cứu [6] giải quyết các kỹ thuật phân nhóm để khai phá mẫu truy cập như dự đoán đường dẫn, nhóm trang, phân nhóm mờ, phân nhóm dựa trên hoạt động của kiến (ant-based clustering), chia cắt đồ thị… Các kết quả thực hiện cho thấy rằng giải thuật phân nhóm mờ cho kết quả chính xác đến 98% để dự đoán mẫu truy cập của người dùng tiềm năng, cao hơn những kỹ thuật khác.

1.3.4. Phương pháp mạng neuron nhân tạo

1.3.4.1. Khái niệm

Mạng lưới neuron nhân tạo [53] là các hệ thống được thúc đẩy bởi phân phối, tính toán song song liên tục trong não bộ cho phép phương pháp này có hiệu quả trong các nhiệm vụ kiểm soát và phân loại sinh học phức tạp. Các mạng lưới thần kinh sinh học thực hiện được điều này có thể được mô hình hóa bằng phương pháp toán học bằng một đồ thị có định hướng, có trọng số cao các nút liên kết với nhau (tế bào thần kinh).

1.3.4.2. Các nghiên cứu liên quan

[80] đã đề xuất một mô hình thông minh để dự đoán các cuộc tấn công lừa đảo dựa trên mạng thần kinh nhân tạo đặc biệt là các mạng thần kinh tự cấu trúc. Tương tự, nghiên cứu [51] giới thiệu một phương pháp để phân loại Bộ định vị tài nguyên thống nhất (URL) thành URL lừa đảo hoặc URL không lừa đảo. Mạng thần kinh nhân tạo đã được đào tạo bằng cách sử dụng tối ưu hóa dòng hạt (PSO) để phân loại URL để cải thiện hiệu suất của mạng neuron nhân tạo.

1.3.4.3. Ưu điểm và hạn chế

- Ưu điểm:

Mạng neuron hoạt động tương tự như não bộ thần kinh của con người, được thu thập các tri thức qua kinh nghiệm thông qua các quá trình huấn luyện và cũng như bộ não, phương pháp này có khả năng lưu trữ thông tin và sử dụng những tri thức trong việc dự đoán các dữ liệu ẩn.

- Hạn chế :

Theo nghiên cứu [46], mạng neuron nhân tạo có một hạn chế là kỹ thuật này được xây dựng dựa trên việc sử dụng một phần thông tin có trong các chuỗi huấn luyện. Vì vậy, phương pháp này đã không sử dụng tất cả các thông tin có trong các

chuỗi đào tạo để thực hiện dự đoán và điều này có thể làm giảm rất nhiều về tính chính xác cho dự đoán.

1.3.5. Các phương pháp phối hợp các phương pháp phổ biến

Các kỹ thuật đơn lẻ sử dụng luật kết hợp, Markov … chỉ sử dụng trên cấu trúc Web đơn giản và có hạn chế về không gian lưu trữ và thời gian thực thi (mô hình luật kết hợp) và phát sinh nhiều trạng thái (các mô hình chuỗi Markov). Bên cạnh đó, ngoài mô hình chuỗi Markov có thứ tự cao, các mô hình chuỗi Markov khác và mô hình ARM không xét đến toàn bộ hành vi của một người dùng trong một phiên truy cập [50]. Các nghiên cứu cho thấy để dự đoán truy cập Web cần kết hợp nhiều phương pháp khác nhau hoặc bổ sung thêm các yếu tố bổ sung để dự đoán tốt hơn.

1.3.5.1. Các công trình liên quan

Nghiên cứu [50] đã giới thiệu một phương pháp xem xét thông tin của trang truy cập và thời gian người dùng viếng thăm các liên kết. Họ đã phân nhóm các phiên truy cập của người dùng dựa vào độ tương tự pair-wise và biểu diễn các kết quả phân nhóm bằng một cây gọi là cây click-tream. Người dùng mới sẽ được đưa vào nhóm theo độ tương tự. Mô hình này cũng có thể được dùng cho hệ khuyến nghị.

Nghiên cứu [109] giới thiệu một phương pháp để dự đoán truy cập Web dùng mô hình chuỗi Markov và tính toán độ tương tự Page. Nghiên cứu đã sử dụng mô hình chuỗi Markov và PageRank để dự đoán truy cập Web. Bài báo đưa ra hai tiếp cận OSim (2nd-Markov kết hợp với mô hình tính toán độ phổ biến và tương tự dùng PageRank) và KSim (1^{st}-Markov dùng cho kết hợp với mô hình tính toán độ phổ biến và tương tự dùng PageRank). Kết quả thu được là OSim có độ chính xác dự đoán cao hơn KSim tuy độ phức tạp về thời gian cao hơn.

Nghiên cứu [61] trình bày một tiếp cận kết hợp khai phá luật kết hợp và các mô hình chuỗi Markov có thứ tự thấp để đạt độ chính xác cao hơn với độ phức tạp

về không gian trạng thái thấp hơn, và các luật kết hợp cũng giúp đạt độ chính xác cao hơn. Các kết quả thực nghiệm của nghiên cứu cho thấy mô hình kết hợp Markov và luật kết hợp làm tăng độ chính xác cho dự đoán truy cập Web.

Nghiên cứu [62] đề xuất một mô hình gọi là là sự kết hợp của các mô hình Clustering, luật kết hợp và Markov. Kết quả họ thu được chứng minh rằng mô hình tích hợp cung cấp dự đoán tốt hơn mỗi mô hình đơn lẻ.

Nghiên cứu [8] kết hợp hai kỹ thuật phân lớp là Markov và Support Vector Machines (SVM) để giải quyết dự đoán bằng cách dùng luật Dempster. Kết quả nghiên cứu cho thấy phương pháp kết hợp này khắc phục giới hạn của mô hình chuỗi Markov (dự đoán dữ liệu ẩn) và SVM (giải quyết số lượng lớn các lớp). Nghiên cứu cũng đã áp dụng trích xuất đặc trưng để tăng hiệu quả cho SVM. Bên cạnh đó, độ chính xác được tăng lên và thời gian dự đoán cũng được thu hẹp khi họ sử dụng tri thức về lãnh vực để giảm số lượng bộ phân lớp (classifiers).

Nghiên cứu [63] đã dùng mô hình gọi là IMAC (Integration of Markov model, Association rules and Clustering) kết hợp các phương pháp Markov thứ tự thấp, Clustering và luật kết hợp. Trong đó Clustering dùng để nhóm các phiên truy cập tương tự của người dùng; Markov được xây dựng trên các phiên truy cập đã được phân nhóm; các luật kết hợp được dùng khi các mô hình chuỗi Markov có thể dự đoán rõ ràng. Kết quả nghiên cứu cho thấy mô hình IMAC hiệu quả hơn các mô hình đơn lẻ và các mô hình kết hợp khác.

Nghiên cứu [7] giới thiệu một mô hình biến thể của all k^{th} Markov tích hợp với mô hình Clustering để dự đoán truy cập Web trên dữ liệu Web Log nhằm làm giảm độ phức tạp không gian trạng thái để cung cấp khuyến nghị tốt hơn.

Một nghiên cứu về dự đoán truy cập Web khác được thực hiện trong bài báo [27] đã trình bày mô hình tích hợp 3 phương pháp: Markov thứ tự thấp, Clustering (giải thuật K-means) và luật kết hợp theo lý thuyết Dempster Shafer. Kết quả nghiên cứu cho thấy mô hình này đạt độ chính xác cao hơn từng mô hình đơn lẻ.

Nghiên cứu [29] đề xuất một mô hình dùng kỹ thuật Clustering (K-means và K-mediods, với K thu được từ giải thuật HITS) và khai phá hành vi truy cập của người dùng thông qua mô hình chuỗi Markov. Kết quả dự đoán truy cập Web hiệu quả hơn về bộ nhớ.

Nghiên cứu [100] sử dụng tích hợp phương pháp Markov và phương pháp Clusring, cụ thể là K-means, để dự đoán truy cập Web trong các tập tin lưu lại lịch sử truy cập Web (log files). Nghiên cứu đã chỉ ra phương pháp Markov tích hợp với Clustering hiệu quả hơn phương pháp Markov tích hợp với luật kết hợp về độ chính xác trong dự đoán truy cập Web.

Để tăng độ chính xác cho dự đoán, nghiên cứu kết hợp giữa Markov và giải thuật PageRank cũng được đề xuất [28]. Ngoài ra một phương pháp tích hợp mô hình chuỗi Markov, Cluster và PageRank đã được trình bày trong bài báo [109]. Cụ thể là, kỹ thuật Clustering được thực hiện để phân nhóm dữ liệu, tiếp theo mô hình chuỗi Markov được sử dụng để dự đoán truy cập Web, và để độ chính xác được tốt hơn giải thuật PageRank dự trên độ tương tự của các trang được thực hiện.

Các mô hình kết hợp và các kỹ thuật tương ứng được các nhà nghiên cứu thực hiện từ năm 2015 đến 2018 được trình bày theo ***Bảng 1.2***.

Bảng 1.2 *Các nghiên cứu dự đoán truy cập Web từ năm 2015 đến năm 2018*

Các mô hình kết hợp (lai)	**Các kỹ thuật tương ứng**
Liraki at al. (2015) [76]	Fuzzy Clustering, Weighted Association Rules, and Fuzzy Inference System
P. Kumar at al. (2015) [69]	Association Rules, Rough Set Clustering
Kundra at al. (2015) [70]	Efficient Hierarchical Particle Swarm Optimization Clustering, Scaled Markov Model, Improved Popularity and Similarity Based PageRank Algorithm (IPSPR)

Rathod at al. (2016) [96]	Fuzzy C-Means Clustering algorithms and Markov
Swarnakar et al. (2016) [108]	K-means Clustering, PageRank, 1st Order Markov
B. H. Kumar at al. (2016) [68]	Hierarchical Clustering, higher order Markov
Poornalatha at al. (2017) [94]	Clustering, Markov, Association Rules
Chembath at al. (2017) [20]	Clustering, Markov, Association Rules and Longest Common Sequence
Yao at al. (2017) [114]	Hidden Markov, Association rules
N. V. Patil at al. (2017) [88]	All Kth Markov, Association Rules, Conditional Sequence Base
Gopalakrishnan at al. (2018) [45]	Fuzzy C-means Clustering, Variable Order Markov

1.3.5.2. Ưu điểm, hạn chế và khuyến nghị

- ***Các tiêu chí đánh giá***
  - ✓ ***Độ chính xác dự đoán:*** Mức độ phù hợp của trang Web kế tiếp tìm thấy so với thực tế. Để độ chính xác dự đoán tốt yêu cầu không bị mất thông tin và không bỏ qua các ứng viên tiềm năng, hay các trường hợp hiếm và giải quyết loại bỏ các thông tin không cần thiết.
  - ✓ ***Độ phức tạp thời gian thực thi dự đoán:*** Giải quyết vấn đề xử lý dự đoán các tập dữ liệu lớn, cũng như không gian dự đoán lớn với độ phức tạp thời gian nhỏ, đảm bảo thời gian thực thi nhanh.
- ***Ưu điểm:***
  - ✓ Ý tưởng chính của phương pháp gom cụm (Clustering) là để cải thiện hiệu năng và tính linh hoạt của các công việc có tính chất cá nhân

hóa [3; 18; 86; 97; 106]. Các phiên truy cập Web có thể được nhận thông qua việc gom cụm các trang hay người dùng.

- ✓ Các mô hình Markov thường được dùng để nhận biết trang Web kế tiếp mà được truy cập bởi người dùng Web dựa trên chuỗi tuần tự các trang Web truy cập trước đó [16; 31; 59; 64; 103; 120].
- ✓ Các nghiên cứu dựa vào luật kết hợp (Association rule) khám phá các luật kết hợp trên các kết quả dữ liệu nhật ký truy cập của người dùng để tìm nhóm các trang Web mà được truy cập cùng nhau [64].
- ✓ Sự tích hợp các tiếp cận khác nhau đã giảm các hạn chế của từng phương pháp cho nhau đã làm tăng hiệu quả truy cập Web, đặc biệt là về phương diện độ chính xác.
- ✓ Nhiều nghiên cứu đã tận dụng thế mạnh của khai phá dữ liệu lịch sử truy cập của người dùng dự đoán truy cập Web. Đây là một chủ đề rất quan trọng trong khai khá dữ liệu và được nhiều nhà nghiên cứu quan tâm.

- ***Hạn chế:***
  - ✓ Các phương pháp khai phá Association Rules rất tốn chi phí thời gian khi xử lý các mẫu có số lượng lớn và dài và được xây dựng trên mô hình không hỗ trợ dự đoán nên trong quá trình dự đoán, thông tin đã bị hao hụt do đó làm giảm đi độ chính xác dự đoán truy cập Web.
  - ✓ Phương pháp phân nhóm cũng là phương pháp dự đoán làm mất thông tin do xây dựng trên mô hình không hỗ trợ dự đoán [46].
  - ✓ Phương pháp quan tâm đến thời gian truy cập của mỗi liên kết tuy quan trọng, nhưng rất khó xác định là người truy cập có thực sự đang xem liên kết đó hay không hay làm việc gì khác không liên quan.
- ***Các khuyến nghị:***

- ✓ Tìm hiểu các phương pháp dự đoán truy cập Web tốt hơn để nâng cao độ chính xác và cải thiện hiệu năng thời gian.
- ✓ Nghiên cứu kết hợp nhiều phương pháp để làm tăng hiệu quả dự đoán.
- ✓ Xem xét thông tin về mối liên hệ giữa các truy cập Web cũng cần được xem xét như thứ tự thời gian giữa các truy cập, tầm ảnh hưởng, độ quan trọng của mỗi liên kết trên Website.

1.4. Phương pháp dự đoán chuỗi dữ liệu tuần tự

Cho một tập hợp các chuỗi tuần tự huấn luyện, vấn đề của dự đoán chuỗi tuần tự là tìm thành phần kế tiếp của một chuỗi tuần tự cho trước bằng cách quan sát các thành phần trước đó [48]. ***Bảng 1.3*** minh họa so sánh về độ chính xác các phương pháp phổ biến về dự đoán chuỗi dữ liệu tuần tự

Bảng 1.3 *So sánh các tiếp cận dự đoán tuần tự [46]*

Tập dữ liệu	**Độ chính xác dự đoán**						
	CPT+	CPT	AKOM	DG	LZ78	PPM	TDAG
BMS	**38.25**	37.90	31.26	36.46	33.46	31.06	6.95
MSNBC	**61.50**	61.64	47.88	55.68	43.64	38.06	31.14
KOSARAK	**37.64**	33.82	20.52	30.82	20.50	23.86	1.06
FIFA	**35.94**	34.56	25.88	24.78	24.64	22.84	7.14

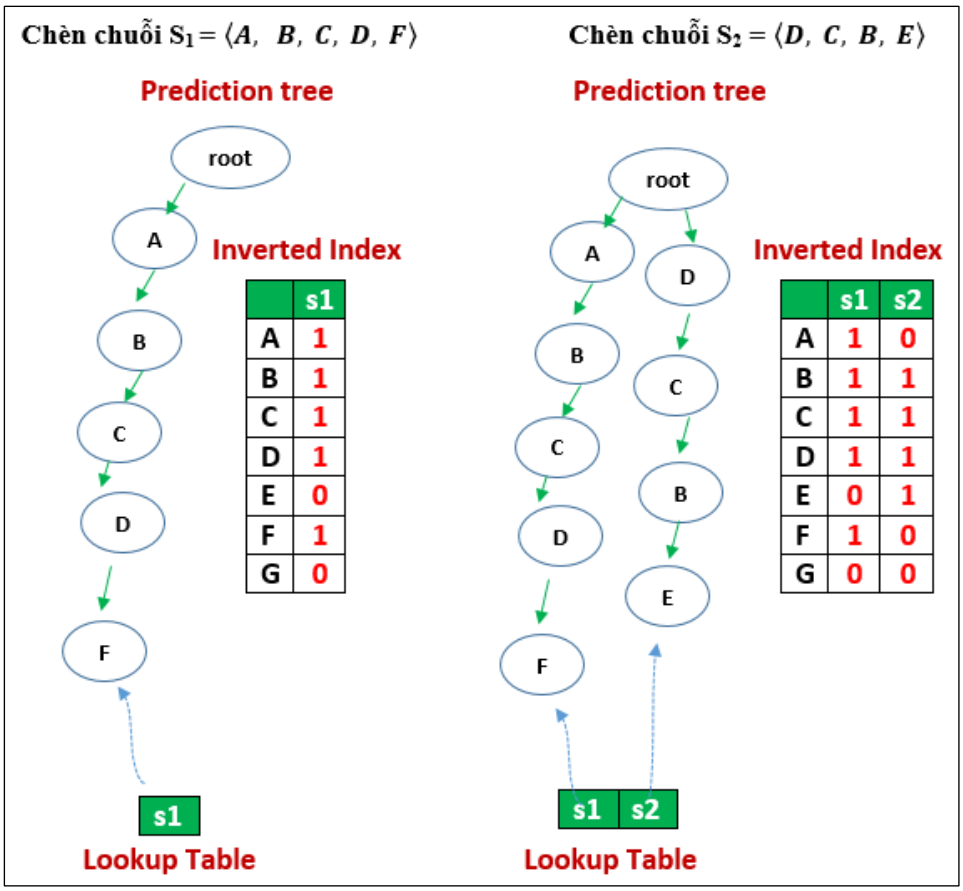
Bảng 1.3 cho thấy phương pháp Markov *(AKOM)* có độ chính xác cao nhất chỉ trên tập dữ liệu *Bible word*. Bên cạnh đó, phương pháp CPT chỉ đạt độ chính xác cao nhất trên tập dữ liệu *MSNBC*. Thống kê trong ***Bảng 1.3*** đã chỉ ra rằng phương pháp CPT+ vượt trội hơn các phương pháp khai phá dữ liệu tuần tự khác khi đạt độ chính xác là cao nhất 5/7 tập dữ liệu quan sát (trong đó hầu hết là các tập dữ liệu truy cập Web như: BMS, MSNBC, FIFA và Kosarak).

Trong các phần tiếp theo, nghiên cứu sinh sẽ giới thiệu về hai phương pháp dự đoán này.

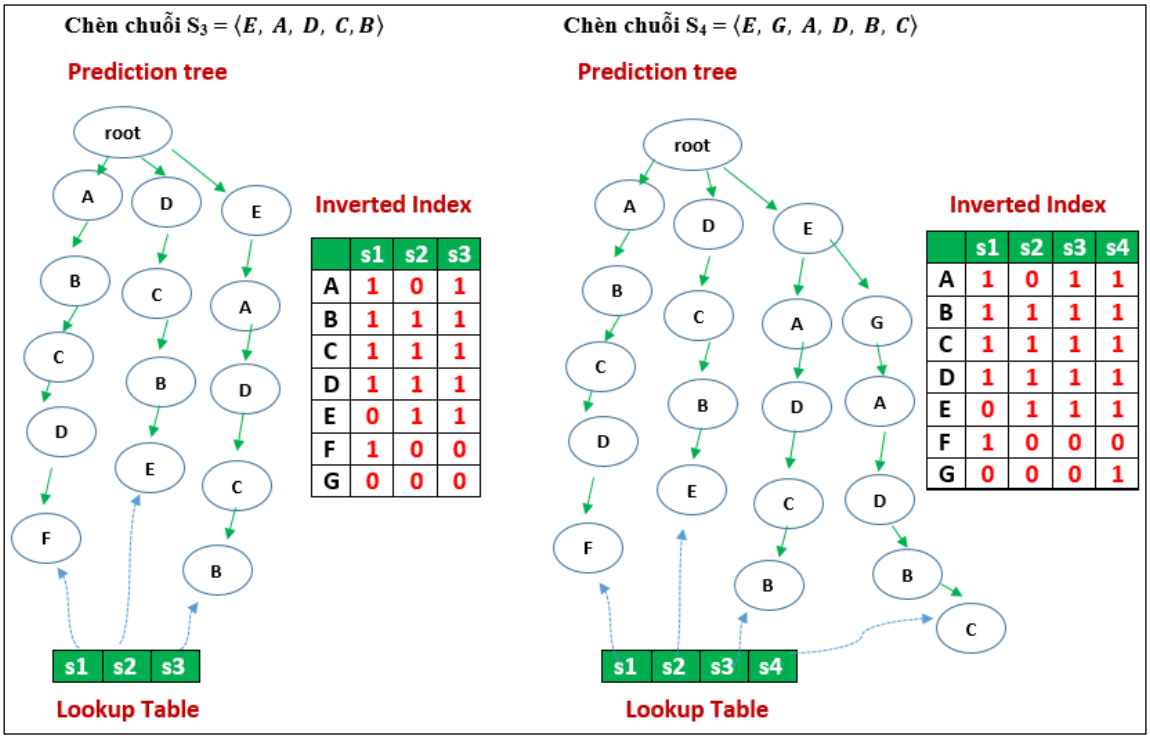
1.4.1. Phương pháp cây dự đoán (Compact Prediction Tree - CPT)

Quá trình huấn luyện của CPT nhập vào một tập các chuỗi tuần tự huấn luyện và tạo ra ba cấu trúc phân biệt: (1) Prediction Tree (PT), (2) Lookup Table (LT) và (3) Inverted Index. Trong suốt quá trình huấn luyện, các chuỗi tuần tự được xem xét từng chuỗi tuần tự để xây dựng dần ba cấu trúc này.

Hình 1.2, Hình 1.3 minh họa việc hình thành của ba cấu trúc bằng cách chèn liên tiếp các chuỗi tuần tự $s_1 = \langle A, B, C, D, F \rangle$, $s_2 = \langle D, C, B, E \rangle$, $s_3 = \langle E, A, D, C, B \rangle$ và $s_4 = \langle E, G, A, D, B, C \rangle$, trong đó bảng chữ cái Z = {A, B, C, D, E, F, G} được sử dụng.



Hình 1.1 Chèn chuỗi s_1 và s_2 vào cây CPT



Hình 1.2 *Chèn chuỗi s_3 và s_4 vào cây CPT*

Một Prefix Tree là một dạng cây tìm kiếm mà ánh xạ một chuỗi thành vài giá trị hay giá trị. Mỗi nút trong cây tiền tố lưu trữ một phần của khóa được xác định bởi vị trí trong cây. Đây là những thuận lợi của việc dùng một cây tiền tố qua một bảng đồ băm. Cây dự đoán (Prediction Tree) là một kiểu của cây tiền tố. Nó chứa tất cả các chuỗi tuần tự huấn luyện. Mỗi nút của cây biểu diễn một phần tử và mỗi chuỗi tuần tự huấn luyện được biểu diễn bởi một đường dẫn bắt đầu từ gốc của cây và kết thúc bằng một nút trong hay một nút lá. Như một cây tiền tố, cây dự đoán là một biểu diễn tinh gọn của các chuỗi tuần tự huấn luyện. Các chuỗi tuần tự chia sẻ một đường dẫn chung trong cây. Lookup Table là một mảng kết hợp mà cho phép định vị bất cứ chuỗi tuần tự huấn luyện nào trong cây dự đoán với một thời gian truy cập không đổi. Cuối cùng Inverted Index là một tập các vector nhị

phân mà xác định mỗi phần tử i từ bảng chữ cái Z, tập hợp các chuỗi tuần tự chứa phần tử i.

Quá trình dự đoán của CPT dựa trên các cấu trúc dữ liệu đã được đề cập ở trên. Đối với một chuỗi tuần tự s = $\langle i_1, i_2, \dots, i_n \rangle$ gồm n phần tử, hậu tố của s kích cỡ y với $1 \leq y \leq n$ được định nghĩa là $P_y(s) = \langle i_{n-y+1}, i_{n-y+2}, \dots, i_n \rangle$. Dự đoán các phần tử kế tiếp của s được thực hiện bằng cách nhận diện các chuỗi tuần tự tương tự như *Py(s)*, nghĩa là các chuỗi tuần tự chứa tất cả phần tử trong *Py(s)* trong bất kỳ thứ tự nào. Chiều dài hậu tố là một tham số tương tự như thứ tự mô hình của All-k-order Markov và DG. Nhận dạng giá trị tối ưu được thực hiện theo kinh nghiệm bắt đầu với chiều dài bằng 1. CPT dùng kết quả của mỗi chuỗi tuần tự tương tự với *s* để thực hiện dự đoán. Đặt $u = \langle j_1, j_2, \dots, j_m \rangle$ là một chuỗi tuần tự tương tự với *s*. Mệnh đề kết quả của *u* đối với *s* là chuỗi tuần tự con dài nhất $\langle j_v, j_{v+1}, \dots, j_m \rangle$ của *u* sao cho $\bigcup_{k=1}^{v-1}\{j_k\} \subseteq P_y(s)$ và $1 \leq v \leq m$. Mỗi phần tử được tìm thấy trong mệnh đề kết quả của một chuỗi tuần tự tương tự nhau của *s* được lưu trong một cấu trúc dữ liệu được gọi là *Count Table* (CT). *Count Table* lưu độ hỗ trợ (tần số) của mỗi phần tử này, mà là một ước lượng của $P(e|P_y(S))$. CPT trả về phần tử (các phần tử) được hỗ trợ tốt nhất trong CT vì những dự đoán của nó.

Ví dụ: Cho chuỗi s = $\langle A, D \rangle$ có 2 phần tử và cơ sở dữ liệu tuần tự gồm 4 chuỗi dữ liệu tuần tự như sau:

s_1: $\langle A, B, C, D, F \rangle$

s_2: $\langle D, C, B, E \rangle$

s_3: $\langle E, A, D, C, B \rangle$

s_4: $\langle E, G, A, D, B, C \rangle$

Tìm phần tử kế tiếp của chuỗi s?

- ✓ Các chuỗi tuần tự tương tự với s là s_1, s_3, s_4
- ✓ Mệnh đề kết quả của s_1 đối với s là $\langle F \rangle$

- ✓ Mệnh đề kết quả của s_3 đối với s là $\langle C, B \rangle$
- ✓ Mệnh đề kết quả của s_4 đối với s là $\langle B, C \rangle$
- ✓ Hậu tố s kích cỡ y: $P_y(s)$ = {C, B, F}
- ✓ Ta có:
  - P({C}|{C,B,F}) = 2/3 = 0.667
  - P({B}|{C,B,F}) = 2/3 = 0.667
  - P({F}|{C,B,F}) = 1/3 = 0.333

Như vậy, có thể kết luận rằng phần tử kế tiếp của chuỗi s = $\langle A, D \rangle$ cần dự đoán khả dĩ là B hoặc C (với điểm cao nhất trong Count Table là 0.667).

- ***Ưu điểm:*** Mô hình dự đoán chuỗi dự liệu tuần tự CPT có ưu thế về độ chính xác so với những tiếp cận khác như khai phá luật kết hợp, khai phá luật liên tiếp, các mô hình phát triển theo Markov.
- ***Hạn chế:*** Theo [46], CPT có thời gian thực thi còn chậm hơn một số giải thuật dự đoán chuỗi tuần tự khác. Do đó cần một tiếp cận cải tiến hơn để giải quyết hạn chế này. Phần tiếp theo sẽ mô tả chi tiết về một cải tiến của CPT.

1.4.2. Phương pháp cây dự đoán cải tiến (Compact Prediction Tree plus - CPT+)

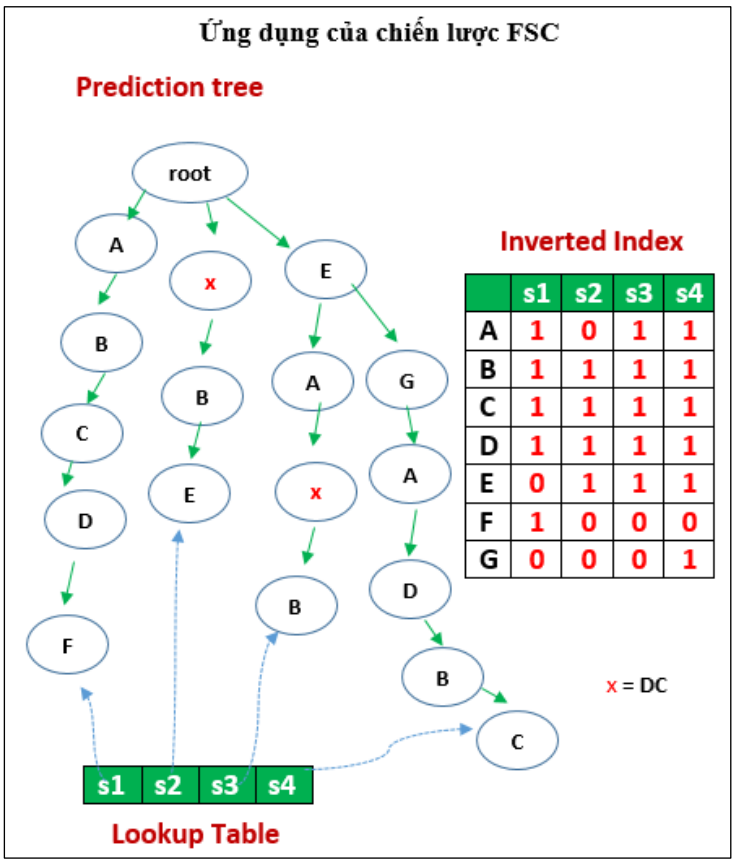
CPT được trình bày như là một trong những mô hình dự đoán chuỗi tuần tự chính xác nhất nhưng độ phức tạp về không gian cao của nó làm cho CPT không thích hợp cho các ứng dụng mà số lượng các chuỗi tuần tự rất lớn. Cây dự đoán CPT là một cấu trúc dữ liệu lớn nhất và chiếm phần lớn về độ phức tạp về không gian của nó.

CPT+ [46] là một biến thể cải tiến từ giải thuật CPT. Đây là một mô hình dự đoán dùng giải pháp nén các chuỗi tuần tự không làm mất mát thông tin bằng cách khai thác các độ tương tự giữa các chuỗi tuần tự con. Độ chính xác của CPT cao hơn nhiều so với các mô hình hiện tại như PPM, DG, AKOM trên các tập dữ liệu thực khác nhau nhưng thời gian dự đoán còn chậm hơn các mô hình này. Một

chiến lược hiệu quả để làm giảm thời gian dự đoán là truy xuất ít thông tin nhất nếu có thể khi dự đoán để tăng tốc độ dự đoán nhưng cũng chọn lọc thông tin cẩn thận để tránh làm giảm độ chính xác. Để giải quyết vấn đề này, một giải thuật cải tiến hơn được xây dựng là CPT+. Theo [46], chi tiết của mô hình CPT+ được cải tiến từ CPT theo các chiến lược sau:

Chiến lược 1: Frequent Subsequence Compression (FSC)

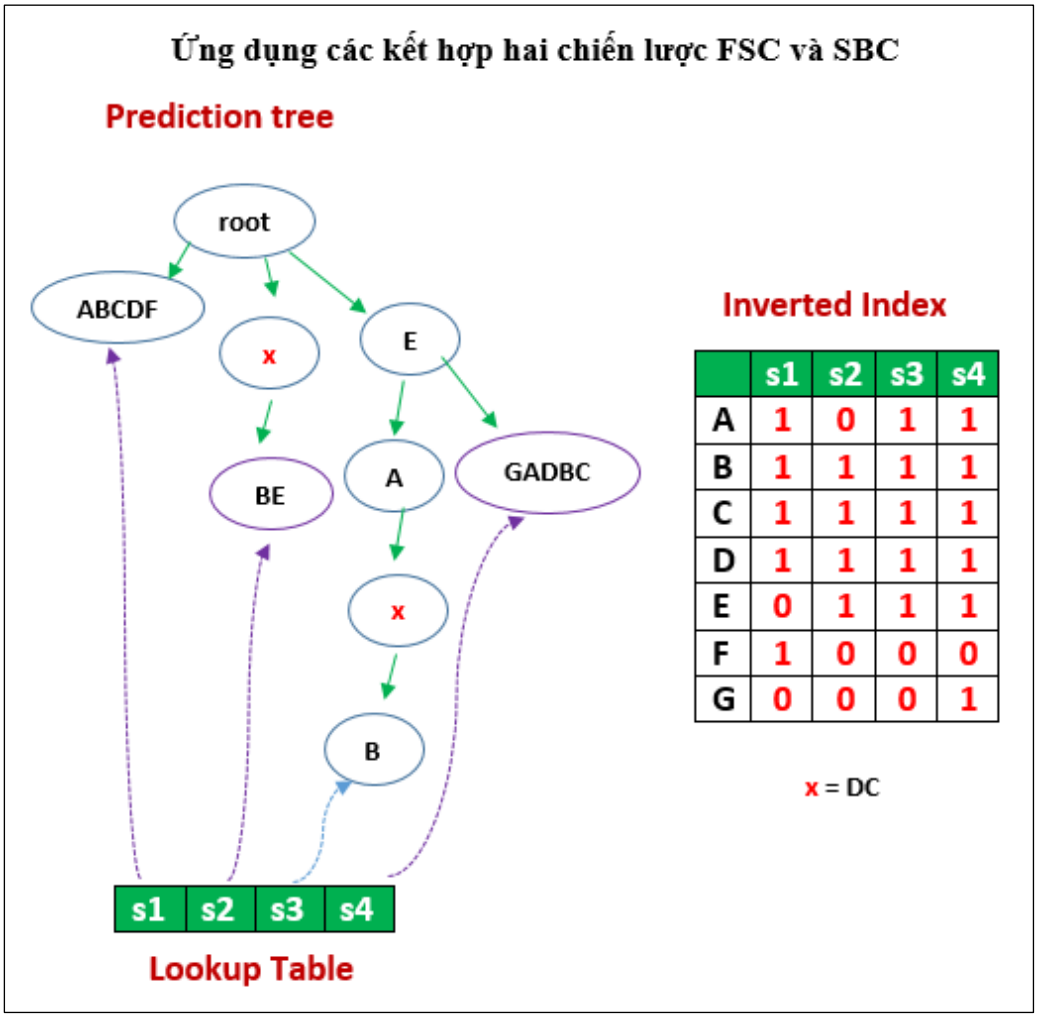
Chiến lược Frequent Subsequence Compression (FSC) ảnh hưởng đến hình dáng của cây dự đoán bằng cách làm giảm độ cao và số lượng nút của nó. Đối với quá trình dự đoán, FSC chỉ ảnh hưởng đến thời gian thực thi. Chi phí bổ sung là sự giảm nén cây dự đoán một cách tự động. ***Hình 1.4*** minh họa cây dự đoán thu được bằng cách áp dụng các chiến lược FSC, các nhánh có chứa DC được thay thế bằng x để giảm chiều cao của cây dự đoán.



Hình 1.3 *Minh họa chiến lược FSC*

Chiến lược 2: Simple Branches Compression (SBC)

Chiến lược SBC bao gồm việc thay thế mỗi nhánh đơn giản bởi một nút biểu diễn toàn bộ nhánh. Chẳng hạn, phần (2) của ***Hình 1.5*** minh họa cây dự đoán thu được bằng cách áp dụng các chiến lược FSC và SBC cho ví dụ đang thực hiện. Chiến lược SBC đã thay thế lần lượt các nhánh đơn giản bởi các nút đơn ABCDF, xBE, AxB, GADBC, với x = DC. Việc nhận biết và thay thế các nhánh đơn giản được thực hiện băng cách duyệt cây dự đoán từ các nút lá dùng Inverted Index. Chỉ có các nút với một nút đơn được viếng thăm.



	s1	s2	s3	s4
A	1	0	1	1
B	1	1	1	1
C	1	1	1	1
D	1	1	1	1
E	0	1	1	1
F	1	0	0	0
G	0	0	0	1

Hình 1.4 *Minh họa chiến lược FSC và SBC*

Chiến lược 3: Prediction with improved Noise Reduction (PNR)

Chiến lược PNR dựa vào giả thiết là độ nhiễu trong các chuỗi tuần tự huấn luyện chứa các phần tử có một tần suất thấp, trong đó một tần suất của phần tử được định nghĩa như là số lượng các chuỗi tuần tự chứa phần tử đó. Vì lý do này, PNR xóa bỏ các phần tử đơn lẻ mà có một tần xuất thấp trong suốt quá trình dự đoán. Vì định nghĩa của độ nhiễu được dùng trong CPT+ là hạn chế hơn so với độ nhiễu của CPT, một số lượng các chuỗi tuần tự con nhỏ hơn được xem xét. Chiến lược PNR đưa vào tham số hậu tố *Py*(*s*) của một chuỗi tuần tự được dự đoán s, các cấu trúc của CPT và tỷ lệ nhiễu *TB* và một số lượng tối thiểu các cập nhật, *MBR,* được thực hiện trên bảng đếm (CT) để thực hiện một dự đoán. PNR là một thủ tục đệ quy, để thực hiện một dự đoán, chúng ra cần PNR xem xét một số lượng tối thiểu các chuỗi tuần tự con được lấy từ *Py*(*s*). Đầu tiên PNR xóa bỏ sự nhiễu từ mỗi chuỗi tuần tự con. Sau đó, bảng đếm CT được cập nhật sử dụng những chuỗi con tuần tự này. Khi số lượng tối thiểu các cập nhật được đạt tới, một dự đoán được thực hiện giống như trong CPT dùng bảng đếm CT. Chiến lược PNR là một sự mở rộng của chiến lược giảm nhiễu được dùng bởi CPT. Ba đóng góp chính được mang đến bởi PNR là cần một số lượng tối thiểu các cập nhật trên bảng đếm CT để thực hiện một dự đoán, xác định nhiễu dựa trên tần số của các phần tử và xác định sự nhiễu tương ứng với độ dài chuỗi tuần tự.

1.4.3. Ưu điểm và hạn chế của phương pháp cây dự đoán cải tiến (CPT+)

- ***Ưu điểm:*** Mô hình dự đoán chuỗi dự liệu tuần tự CPT+ có ưu thế về độ chính xác và thời gian so với những tiếp cận khác như khai phá luật kết hợp, khai phá luật liên tiếp, các mô hình phát triển theo Markov, CPT. Điều này được lý giải là vì CPT+ được xây dựng trên một mô hình mà sự mất thông tin có thể được quản lý. Hơn nữa, CPT+ sử dụng tất cả thông tin liên quan để thực hiện dự đoán và thời gian xử lý dự đoán rất nhanh vì CPT+ áp dụng hai chiến lược nén rất hiệu quả là Compressing Frequent Substrings và

Compressing Simple Branches. Hơn nữa, để nâng cao độ chính xác dự đoán, CPT+ cũng áp dụng chiến lược Improved Noise Reduction để loại bỏ các dữ liệu không có ý nghĩa cho dự đoán.

- ***Hạn chế:*** Để dự đoán truy cập Web, tương tự như các mô hình dự đoán chuỗi tuần tự khác, phương pháp cây dự đoán cải tiến (CPT+) vẫn cần giải quyết các vấn đề về:
  - ✓ Thời gian thực thi dự đoán còn chậm nếu không gian dự đoán lớn [46; 48] . Vì thế cần đề xuất các giải pháp để làm tăng tốc độ thời gian dự đoán mà độ chính xác vẫn bảo toàn. Ví dụ như xem xét tính chất của chuỗi tuần tự truy cập Web cần dự đoán truy cập kế tiếp với các cơ sở dữ liệu tuần tự truy cập Web.
  - ✓ Nâng cao độ chính xác cho dự đoán: Xem xét các mối quan hệ, tương tác giữa các trang với nhau để đưa ra các giải pháp để nâng cao hiệu quả về chính xác cho dự đoán truy cập Web. Chẳng hạn như độ tin cậy, tầm ảnh hưởng của các trang Web bằng cách giải thuật PageRank.

1.4.4. Tổng hợp so sánh các phương pháp dự đoán chuỗi dữ liệu tuần tự

Theo nghiên cứu [46], dữ liệu trên ***Bảng 1.4*** cho thấy trên tập dữ liệu BMS (gồm có 15, 806 chuỗi dữ liệu tuần tự), phương pháp CPT+ có độ chính xác vượt trội hơn những phương pháp phổ biến thường dùng để dự đoán chuỗi tuần tự khác như CPT, DG, PPM và AKOM.

Bảng 1.4 *Bảng so sánh độ chính xác các phương pháp dự đoán chuỗi dữ liệu tuần tự*

Tập dữ liệu truy cập Web	**Độ chính xác dự đoán** *(%)*				
	CPT+	**CPT**	**DG**	**PPM**	**AKOM**

BMS (15,806)	**38.25**	37.90	36.46	31.06	31.26
FIFA (573, 060)	**35.94**	34.56	24.78	22.84	25.88
KOSARAK (638,811)	**37.64**	33.82	30.82	23.86	20.52

Về thời gian thực thi dự đoán, các kết quả thực nghiệm trong nghiên cứu [48] và [46] được tổng hợp trong ***Bảng 1.5*** cho thấy trên tập các dữ liệu BMS, FIFA, KOSARAK phương pháp CPT+ có thời gian thực thi dự đoán nhanh xấp xỉ 4.5 lần phương pháp CPT, tuy nhiên thời gian thực thi dự đoán của phương pháp CPT+ là chậm hơn so với các phương pháp DG, PPM và AKOM trên tập dữ liệu BMS; với tập dữ liệu FIFA, phương pháp CPT+ có thời gian thực thi dự đoán chỉ chậm hơn so với phương pháp PPM; trên tập dữ liệu KOSARAK thời gian thực thi dự đoán của phương pháp CPT+ là chậm hơn so với các phương pháp DG, PPM và AKOM.

Bảng 1.5 *Bảng so sánh thời gian thực thi các mô hình dự đoán*

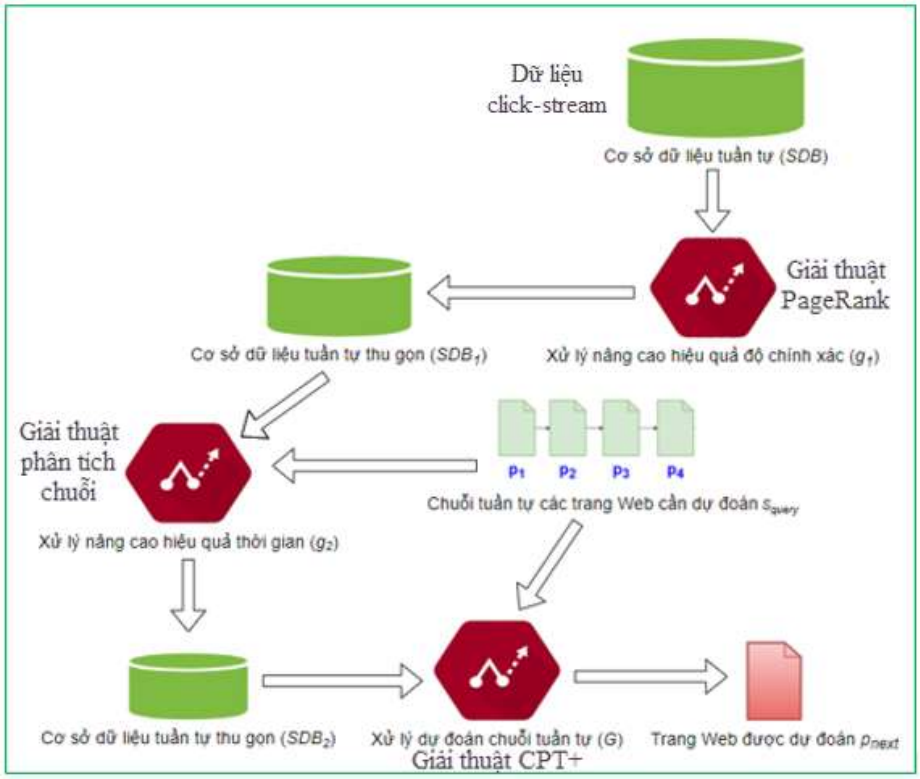
Tập dữ liệu truy cập Web	**Thời gian thực thi dự đoán** *(seconds)*				
	CPT+	**CPT**	**DG**	**PPM**	**AKOM**
BMS (15,806)	0.078	0.352	0.004	0.001	0.004
FIFA (573, 060)	0.032	0.146	0.301	0.006	0.085
KOSARAK (638,811)	0.341	1.533	0.042	0.018	0.011

Tóm lại, mặc dù có nhiều ưu điểm so với tiếp cận phổ biến trong dự đoán chuỗi dữ liệu tuần tự, phương pháp CPT+ cũng còn một số hạn chế sau: (1) Thời

gian xử lý chậm nếu cơ sở dữ liệu tuần tự chứa nhiều chuỗi tuần tự có số phần tử truy cập lớn và kích cỡ của cơ sở dữ liệu tuần tự càng lớn thì càng ảnh hưởng đến thời gian thực thi dự đoán; (2) Chưa xử lý triệt để dữ liệu dư thừa do đó độ chính xác còn bị ảnh hưởng.

1.5. Đề xuất mô hình dự đoán hành vi truy cập Web

Như đã được định nghĩa ở phần 1.3, s_{query} là một chuỗi tuần tự gồm có k trang Web $p_1, p_2, \dots p_k$ được truy cập với $s_{query} = \langle p_1, p_2, \dots p_k \rangle$. Để giải quyết bài toán dự đoán hành vi truy cập Web, cụ thể là tìm trang Web p_{next} là trang kết quả dự đoán, luận án đề xuất mô hình như sau: Mô hình khai phá dữ liệu cho dự đoán truy cập Web theo hướng kết hợp nâng cao độ chính xác và nâng cao hiệu quả về thời gian. Cụ thể, luận án đề xuất dự đoán truy cập Web bằng cách kết hợp các giải pháp: Xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web, nâng cao độ chính xác cho dự đoán truy cập Web (**Chương 3**) và nâng cao hiệu quả thời gian cho dự đoán truy cập Web (**Chương 4**). Mô hình được thể hiện một cách trực quan theo ***Hình 1.6.***



Hình 1.5 *Mô hình khai phá dữ liệu cho dự đoán truy cập Web kết hợp nâng cao độ chính xác và nâng cao hiệu quả về thời gian*

Diễn giải mô hình :

Bước 1: Xây dựng cơ sở dữ liệu tuần tự truy cập Web (Chi tiết được trình bày ở **Chương 2**)

$$SDB = f_0\,(L) \qquad (1.11)$$

Trong đó: Cơ sở dữ liệu tuần tự *SDB* là cơ sở dữ liệu được xây dựng theo một hàm xử lý f_0.

Bước 2: Nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web (Chi tiết được trình bày ở **Chương 3**)

$$SDB_1 = g_1\,(SDB) \qquad (1.12)$$

Cơ sở dữ liệu tuần tự SD_1 là cơ sở dữ liệu *SD* được thu gọn bằng giải pháp loại bỏ các chuỗi tuần tự dư thừa bằng cách dùng hàm xử lý g_1, cụ thể là giải thuật PageRank.

Bước 3: Nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web (Chi tiết được trình bày ở **Chương 4**)

Cơ sở dữ liệu tuần tự truy cập Web ở bước này được xác định bằng công thức:

$$SDB_2 = g_2 (s_{query}, SDB_1) \quad (1.13)$$

Trong đó, cơ sở dữ liệu tuần tự truy cập Web SDB_2 là cơ sở dữ liệu tuần tự truy cập Web SDB_1 được thu gọn bằng giải pháp loại bỏ các chuỗi tuần tự dư thừa bằng cách dùng hàm xử lý g_2, cụ thể là giải thuật phân tích và so sánh chuỗi.

Bước 4: Trang Web kết quả dự đoán p_{next} được xác định bằng một hàm xử lý G, cụ thể là CPT+ với dữ liệu đầu vào là chuỗi tuần tự cần dự đoán S_{query} và cơ sở dữ liệu tuần tự đã được thu gọn SBD_2.

$$p_{next} = G(s_{query}, SDB_2) \quad (1.14)$$

1.6. Các giải pháp đề xuất

Các nghiên cứu trên cho thấy công việc dự đoán truy cập Web cần kết hợp nhiều phương pháp để đảm bảo độ chính xác nhưng thời gian thực hiện cũng cần được xem xét. Do đó việc nâng cao hiệu quả về thời gian, nâng cao hiệu quả về độ chính xác cho dự đoán hành vi truy cập Web là rất cần thiết. Nghiên cứu sinh giới thiệu mô hình kết hợp các giải pháp để dự đoán truy cập Web hiệu quả hơn:

(i) Đề xuất phương pháp thiết kế và xây dựng cơ sở dữ liệu tuần tự thu gọn từ Web Log để thuận tiện cho dự đoán truy cập Web.

(ii) Đề xuất giải pháp nâng cao hiệu quả về thời gian nhưng vẫn đảm bảo độ chính xác cho dự đoán hành vi truy cập Web: Giảm không gian dự đoán để tăng thời gian thực thi dự đoán truy cập Web bằng cách kết hợp tích hợp PageRank vào phương pháp CPT+.

(iii) Đề xuất giải pháp nâng cao hiệu quả về thời gian truy cập Web nhưng phải đảm bảo độ chính xác cho dự đoán: Giảm thời gian thực thi bằng cách giảm không gian dự đoán trên cơ sở tận dụng sức mạnh về dự đoán chuỗi tuần tự của CPT+ kết hợp với giải pháp phân tích chuỗi để

loại bỏ các chuỗi tuần tự dư thừa, không có nghĩa trong quá trình dự đoán truy cập Web.

(iv) Đề xuất giải pháp kết hợp nâng cao độ chính xác và hiệu quả về thời gian truy cập Web.

1.7. Kết luận chương 1

Để dự đoán truy cập Web, các nhà nghiên cứu đã thực hiện các nghiên cứu khác nhau từ các phương pháp độc lập như khai phá luật kết hợp, các mô hình phát triển từ Markov, CPT và CPT+… đến các phương pháp kết hợp các mô hình khác nhau. Trên cơ sở nghiên cứu và phân tích các điểm mạnh và yếu của các phương pháp dự đoán hành vi truy cập Web, luận án đã đề xuất và xuất bản công trình nghiên cứu [CT5]. Các nghiên cứu tiếp theo được nghiên cứu sinh đề xuất như sau:

- ✓ Đề xuất giải pháp xây dựng cơ sở dữ liệu tuần tự để dự đoán truy cập Web. Giải pháp này được trình bày chi tiết trong **Chương 2**.
- ✓ Đề xuất giải pháp nâng cao hiệu quả về độ chính xác cho truy cập Web (tích hợp giải thuật PageRank với CPT+). Giải pháp này được trình bày chi tiết trong **Chương 3**.
- ✓ Đề xuất giải pháp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web bằng phương pháp kết hợp giải thuật CPT+ với phân tích chuỗi dự đoán. Giải pháp này được trình bày chi tiết trong **Chương 4**.
- ✓ Đề xuất giải pháp kết hợp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web và nâng cao độ chính xác cho dự đoán truy cập Web. Giải pháp này được trình bày chi tiết trong **Chương 5**.

CHƯƠNG 2. XÂY DỰNG CƠ SỞ DỮ LIỆU TUẦN TỰ CHO DỰ ĐOÁN TRUY CẬP WEB

2.1. Giới thiệu

Chương 2 trình bày một giải pháp xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web. Cơ sở dữ liệu tuần tự được xây dựng từ các chuỗi tuần tự của tập dữ liệu click-stream hoặc tập dữ liệu được chuẩn hóa từ nhật ký của máy chủ Web. Việc chuẩn hóa dữ liệu từ máy chủ Web là quá trình tiền xử lý để làm sạch và biến đổi dữ liệu để phục vụ cho dự đoán truy cập Web. Bên cạnh dữ liệu clickstream đã được chuẩn hóa vì mang tính chất của dữ liệu tuần tự, nghiên cứu sinh đã đề xuất một giải thuật để chuẩn hóa và xây dựng cơ sở dữ liệu tuần tự cho dữ liệu nhật ký Web. Cơ sở dữ liệu tuần tự click-stream được dùng cho các giải pháp nâng cao hiệu quả về độ chính xác cho dự đoán hành vi truy cập Web (được trình bày ở Chương 3) và nâng cao hiệu quả về thời gian dự đoán hành vi truy cập Web (được trình bày ở Chương 4). Cơ sở dữ liệu tuần tự từ nhật ký Web được dùng cho nâng cao hiệu quả về thời gian dự đoán hành vi truy cập Web (được trình bày ở Chương 4).

2.2. Cơ sở lý luận của giải pháp

Web Log là tập hợp các tập tin nhật ký Web được thu thập từ máy chủ Web. Các tập tin này chứa một khối lượng rất lớn dữ liệu được ghi nhận lại trong toàn bộ quá trình một Website hoạt động. Bên cạnh đó, Web Log cũng chứa nhiều thông tin lỗi, dư thừa, nhiễu thông tin, gây hiểu nhầm và không đầy đủ. Chẳng hạn, sự lặp đi lặp lại các truy cập trên cùng một liên kết của cùng một người dùng tại các thời điểm liên tiếp nhau gây ra sự dư thừa trên Web Log. Điều này thực sự không đem lại lợi ích cho dự đoán. Hơn nữa, các tập tin nhật ký Web cũng chứa những liên kết xấu, thiếu hoặc không cho phép truy cập, hoặc truy cập lỗi. Chẳng hạn như các lỗi liên kết có dạng 3xx, 4xx và 5xx [22]. Vì thế, việc thực hiện tiền xử lý để

loại bỏ những thông tin không cần thiết và không có ý nghĩa để phục vụ dự đoán truy cập Web là công việc rất quan trọng và cần thiết.

Tóm lại, dữ liệu Web Log được dùng cho khai phá dữ liệu tuần tự không phù hợp để sử dụng trực tiếp. Vì vậy, dữ liệu nhật ký web phải được chuyển đổi thành dữ liệu tuần tự và công việc tiền xử lý là rất cần thiết để tránh nhiễu thông tin, các ngoại lệ và các giá trị bị thiếu [115]. Mục đích của tiền xử lý và biến đổi dữ liệu là để có được dữ liệu sạch đáp ứng cho nghiên cứu dự đoán truy cập Web.

Một hạn chế đáng chú ý cần xem xét khi dự đoán truy cập Web trên Web Log là thời gian truy cập rất chậm do khối lượng thu thập dữ liệu trên các tập tin nhật ký là cực kỳ lớn. Thật vậy, những thông tin dư thừa, không có ý nghĩa chứa trong các tập tin nhật ký Web đã làm cho độ phức tạp về thời gian cho dự đoán truy cập Web tăng lên.

Do vậy, việc thu hẹp kích thước của Web Log, thu hẹp phạm vi, không gian dự đoán là công việc rất quan trọng để thời gian dự đoán được giảm xuống đến mức thấp nhất có thể mà vẫn đảm bảo độ lớn và độ chính xác của thông tin truy cập Web cần dự đoán.

2.3. Khái niệm Web Usage Mining

2.3.1. Định nghĩa Web Usage Mining

Theo nghiên cứu [105] Web Mining được chia thành ba loại khác nhau: Web Usage Mining, Web Content Mining and Web Structure Mining. Trong đó, Web Usage Mining là một chủ đề quan trọng và có nhiều ứng dụng trên thực tế.

Web Usage Mining là một ứng dụng của các kỹ thuật khai phá dữ liệu để tìm ra các mẫu truy cập lịch sử thu được từ dữ liệu Web để hiểu và phục vụ tốt hơn nhu cầu của các ứng dụng trên nền tảng Web [104].

Hơn nữa, nghiên cứu [13] chỉ ra rằng Web Usage Mining là một kỹ thuật khai phá Web được dùng để tìm và phân tích các mẫu lịch sử truy cập Web từ dữ

liệu lịch sử Web (còn gọi là các Web Log) hay nói cách khác Web Usage Mining chính là Web Log Mining.

Web Usage Mining gồm ba giai đoạn là tiền xử lý, khai phá mẫu và phân tích mẫu. Trong giai đoạn tiền xử lý của Web Usage Mining, nghiên cứu sinh tiến hành xây dựng cơ sở dữ liệu tuần tự. Việc xây dựng cơ sở dữ liệu này thực chất bao gồm hai giai đoạn chủ yếu là giai đoạn làm sạch dữ liệu và giai đoạn chuyển đổi dữ liệu. Trong đó, giai đoạn làm sạch dữ liệu bao gồm các thao tác sửa dữ liệu xấu, lọc một số dữ liệu không chính xác khỏi bộ dữ liệu và giảm chi tiết không cần thiết của dữ liệu; trong giai đoạn chuyển đổi dữ liệu, dữ liệu được chuyển đổi hoặc hợp nhất để kết quả quá trình khai thác có thể được hiệu quả hơn [42].

2.3.2. Tầm quan trọng của Web Usage Mining

Trong nhiều năm gần đây, rất nhiều nghiên cứu đã được xuất bản để mô tả những bước tiến lớn trong lãnh vực liên quan đến Web Usage Mining [19]. Các kết quả của nghiên cứu này trở nên quan trọng cho rất nhiều ứng dụng như hỗ trợ ra quyết định kinh doanh và tiếp thị, nghiên cứu khả năng truy cập Web, phân tích lưu lượng mạng. Bên cạnh đó, tri thức thu được từ các mẫu truy cập lịch sử Web có thể ứng dụng trực tiếp để quản lý hiệu quả các hoạt động liên quan đến thương mại điện tử, dịch vụ điện tử, giáo dục điện tử... Chẳng hạn, trong thương mại điện tử, phân tích thông tin lịch sử truy cập Web có thể có ích để thu hút khách hàng mới, duy trì khách hàng thân thiết, cải thiện tiếp thị bán hàng, tăng cường hiệu quả các chiến dịch quảng cáo... Các ứng dụng thông thường khác trong Web Usage Mining là các ứng dụng mà mang lại lợi ích từ các kỹ thuật mô hình hóa người dùng như thiết kế các Website tự thích ứng và các hệ khuyến nghị. Thực vậy, Web Usage Mining là một trong những tiếp cận được sử dụng nhiều nhất cho sự phát triển của các hệ thống Website, và được mô tả bởi rất nhiều các nghiên cứu được xuất bản về chủ đề này [2; 24; 90].

2.3.3. Khái niệm cơ sở dữ liệu Web Log

2.3.3.1 Định nghĩa cơ sở dữ liệu Web Log

Các máy chủ Web (Web server) đăng ký một Web log đối với mỗi truy cập đơn lẻ mà chúng nhận được, trong đó các phần quan trọng của thông tin về truy cập được ghi nhận bao gồm URL truy cập, địa chỉ IP từ máy khách (Web client) và thời gian truy cập. [89]

Các tập tin Web log được chia thành nhiều phần nhỏ cho mục đích khai phá dữ liệu nào đó. Để thu được các phần của các Web log, kỹ thuật tiền xử lý sẽ được áp dụng. Mỗi phần của Web log là một chuỗi tuần tự các sự kiện từ một người dùng hay phiên truy cập theo thứ tự thời gian tăng dần, chẳng hạn sự kiện nào đến sớm hơn xảy ra trước sự kiện đến trễ hơn. [89] định nghĩa thành phần Web log (hay còn gọi là chuỗi tuần tự truy cập Web) như sau:

2.3.3.2 Cấu trúc và nội dung Web Log

Cấu trúc và nội dung của Web Log phụ thuộc vào máy chủ tạo ra các Web Log đó. Theo nghiên cứu [19], đa số các máy chủ Web hỗ trợ dưới dạng tùy chọn mặc định, định dạng tập tin nhật ký chung [77] (CLF). CLF còn được gọi là Định dạng Nhật ký Chung NCSA, là định dạng tệp văn bản được tiêu chuẩn hóa được sử dụng bởi các máy chủ web khi tạo tệp nhật ký máy chủ. Mỗi dòng trong một tập tin lưu trữ trong CLF thường bao gồm thông tin như địa chỉ IP của người dùng được kết nối, thời gian truy cập (ngày và giờ truy cập), URL của trang được yêu cầu, giao thức yêu cầu, mã cho biết trạng thái của yêu cầu. Các ví dụ khác về định dạng của tập tin nhật ký được biểu thị bằng Định dạng Nhật ký mở rộng (W3C) được hỗ trợ bởi các máy chủ Web là Apache và Netscape, và định dạng W3SVC tương tự được hỗ trợ bởi IIS (Microsoft Internet Information Server). Các định dạng như vậy được đặc trưng bởi việc bao gồm thông tin bổ sung về các yêu cầu của người dùng, như địa chỉ của URL giới thiệu đến trang được yêu cầu, tên và

phiên bản trình duyệt được người dùng sử dụng để điều hướng, hệ điều hành của máy chủ. Chi tiết các thuộc tính của W3SVC được mô tả như sau:

- ✓ *date:* Ngày thời gian truy cập.
- ✓ *time:* Thời gian truy cập (theo giờ UTC).
- ✓ *s-sitetime:* Tên dịch vụ Internet và mã số đối tượng đang thực thi trên máy khách.
- ✓ *s-computername:* Tên của máy chủ tạo ra tập tin log.
- ✓ *s-ip:* Địa chỉ IP của máy chủ tạo ra tập tin log.
- ✓ *cs-method:* Phương thức yêu cầu truy cập, ví dụ POST, GET
- ✓ *cs-uri-stem:* Liên kết (URL) của *cs-method*, ví dụ portal.ptit.edu.vn/tap-chi/tap-chi-khoa-hoc-cong-nghe-ttt
- ✓ *cs-uri-query:* Truy vấn (nếu có) mà máy khách đang cố gắng thực thi. Một truy vấn URI (Universal Resource Identifier) chỉ cần thiết cho các trang Web động.
- ✓ *s-port:* Mã số cổng của máy chủ cấu hình cho dịch vụ.
- ✓ *cs-username:* Tên của người dùng đã được xác thực mà truy cập vào máy chủ. Những người dùng ẩn danh được xác định bằng một dấu gạch nối "-" (hyphen).
- ✓ *c-ip:* Địa chỉ IP của máy khách mà thực hiện yêu cầu truy cập.
- ✓ *cs-version:* Phiên bản giao thức HTTP mà máy khách đã sử dụng.
- ✓ *cs(User-Agent):* Loại trình duyệt mà máy khách đã sử dụng.
- ✓ *cs(Cookie):* Nội dung của cookie được gửi hay được nhận (nếu có)
- ✓ *cs(Referrer):* Liên kết mà người dùng viếng thăm lần cuối cùng và cung cấp một liên kết đến trang hiện tại.
- ✓ *cs-host:* Tên dịch vụ máy chủ (nếu có).
- ✓ *sc-status:* Mã số giao thức HTTP
- ✓ *sc-substatus:* Mã lỗi của substatus

- ✓ *sc-win32-status:* Mã số tình trạng Windows
- ✓ *sc-bytes:* Số lượng byte được gửi bởi máy chủ
- ✓ *cs-bytes:* Số lượng byte được nhận và thực thi bởi máy chủ
- ✓ *time-taken:* Khoảng thời gian truy cập được thực hiện (tính bằng phần ngàn giây, milliseconds).

Một ví dụ minh họa cho cấu trúc và nội dung Web Log như sau: Để truy vấn thông tin của cơ sở dữ liệu Web log (được lưu trong tập tin Weblog trên máy chủ của Website), trước hết, chúng ta tìm hiểu nội dung của tập tin Web log từ Website chuyên về đặt phòng phục vụ cho du lịch www.villazest.co.za. Một phần thông tin truy cập Web của một số người dùng trong một tập tin Web log được thu thập từ Website www.villazest.co.za được mô tả chi tiết trong ***Bảng 2.1***.

Bảng 2.1 *Minh họa thông tin truy cập của người dùng trên tập tin Web Log*

Địa chỉ IP	**URL truy cập**	**Thời điểm truy cập**
212.198.132.111	cape-town/boutique-hotel/gallery.html	07/Feb/2015 21:45:52
212.198.132.111	cape-town/boutique-hotel/rooms.html	07/Feb/2015 21:46:03
197.86.202.98	cape-town/boutique-hotel/specials.html	09/Feb/2015 04:28:26
197.86.202.98	cape-town/boutique-hotel/rates-bookings.html	09/Feb/2015 04:28:27
197.86.202.98	cape-town/boutique-hotel/contact.html	09/Feb/2015 04:28:27
197.86.202.98	cape-town/boutique-hotel/contact.html	09/Feb/2015 04:28:28

105.233.74.160	cape-town/boutique-hotel/rates-bookings.html	12/Feb/2015 08:37:28
105.233.74.160	cape-town/boutique-hotel/specials.html	12/Feb/2015 08:37:29

Bảng 2.1 cho thấy người dùng Web có địa chỉ IP là *212.198.132.111* truy cập hai trang Web *gallery.html* và *rooms.html* vào hai thời diểm liên tiếp nhau, thời gian chuyển trang khoảng một phút. Đặc biệt, người dùng có địa chỉ IP là *197.86.202.98* đã truy cập liên tiếp vào trang *contact.html* hai lần, khoảng thời gian chuyển giao giữa hai lần truy cập xấp xỉ một giây.

2.3.4. Xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web

2.3.4.1. Mục tiêu

Việc xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web có ý nghĩa rất quan trọng trong khai phá dữ liệu tuần tự vì cơ sở dữ liệu tuần tự được hình thành từ dữ liệu thu thập từ dữ liệu nhật ký Web vốn rất chứa nhiều thông tin dư thừa không cần thiết và gây khó khăn trong việc dự đoán.

Việc sử dụng luôn các mối liên kết giữa các trang Web có sẵn trên Webserver có một số so với việc phải lấy thông tin này từ dữ liệu truy cập của người dùng. Cụ thể, lấy thông tin trên Webserver sẽ có nhiều thông tin hơn so với lấy liên kết truy cập thông thường vì Webserver cung cấp những thông tin: Thời gian truy cập của liên kết, thứ tự truy cập liên kết, địa chỉ IP người dùng truy cập. Những thông tin này rất quan trọng và có ý nghĩa để hỗ trợ cho dự đoán các chuỗi dữ liệu tuần tự phục vụ cho dự đoán truy cập Web.

Một số ứng dụng của cơ sở dữ liệu tuần tự được sử dụng trong trong nhiều nghiên cứu liên quan đến khai phá dữ liệu như khai phá mẫu tuần tự [33; 54; 117], khai phá luật tuần tự [32; 33; 37; 39] và đặc biệt là dự đoán chuỗi tuần tự [23; 46; 48; 72; 84; 93; 121]. Trong phạm vi nghiên cứu này, cơ sở dữ liệu tuần tự được xây

dựng phục vụ cho dự đoán truy cập Web từ dữ liệu thu thập từ nhật ký truy cập Web của người dùng.

2.3.4.2. Dữ liệu

Các cơ sở dữ liệu Web Log được thu thập từ các Website dưới đây:

Website 1: *periwinklelecottages.com*

Website 2: *palmviewsanibel.com*

Website 3: *devqa.robotec.co.il*

Website 4: *inees.org*

Thông tin của các cơ sở dữ liệu Web Log được trình bày như minh họa của ***Bảng 2.3***.

Bảng 2.2 *Thông tin các cơ sở dữ liệu Web Log*

	Website 1	**Website 2**	**Website 3**	**Website 4**
Số lượng Các chuỗi tuần tự	3511237	4217568	2527429	593367
Kích cỡ (MB)	97449	74814	629	119
Số lượng các IP khác nhau	61015	40901	7405	1188
Số lượng các liên kết khác nhau	4267	3535	5467	451

**** Phân tích đặc trưng tập dữ liệu Web Log từ các Website trên :***

+ Dữ liệu Web Log luôn tăng trưởng về mặt khối lượng vì dữ liệu truy cập trên các trang Web không ngừng tăng lên theo thời gian: từng giây, từng phút, từng giờ, ...

+ Tốc độ tăng trưởng của dữ liệu Web Log cũng được tăng lên một cách nhanh chóng vì các truy cập liên tục của người dùng, thông tin truy cập được cập nhật liên tục. Do đó tốc độ thay đổi thông tin là rất nhanh.

2.3.4.3. Phương pháp

Để xây dựng cơ sở dữ liệu tuần tự, các thuộc tính của cơ sở dữ liệu Web Log sau đây được xem xét: IP truy cập của người dùng (*User_IP*), Liên kết truy cập (*Link*), thời điểm truy cập (*Action_Time*). Tùy theo Web Log mà các thuộc tính này có thể được ký hiệu theo quy ước riêng. Hai giai đoạn chính để xây dựng cơ sở dữ liệu từ cơ sở dữ liệu Web Log được trình bày như dưới đây.

Giai đoạn 1: Sắp xếp cơ sở dữ liệu Web Log theo từng User_IP

Biểu diễn mỗi User_IP sao cho trình tự thời gian truy cập của người dùng của tăng dần. ***Bảng 2.2*** minh họa một số mẫu tin của một cơ sở dữ liệu Web Log đã được sắp xếp tăng dần theo thời gian truy cập của từng User_IP.

Bảng 2.3 *Minh họa một phần cơ sở dữ liệu Web Log*

User_IP	**Link**	**Action_Time**
176.9.34.172	Link_visited_1	17:14:11, 12-May-2017
176.9.34.172	Link_visited_4	05:17:21, 18-May-2017
176.9.34.172	Link_visited_5	12:14:12, 20-May-2017
182.92.18.13	Link_visited_3	11:14:16, 12-Apr-2017
182.92.18.13	Link_visited_2	18:04:23, 14-Apr-2017
182.92.18.13	Link_visited_5	21:12:28, 15-Apr-2016
182.92.18.13	Link_visited_6	09:14:23, 17-Apr-2017
170.23.11.67	Link_visited_2	17:14:19, 05-Jun-2017
170.23.11.67	Link_visited_3	06:17:25, 17-Jun-2017
170.23.11.67	Link_visited_1	08:14:06, 23-Jun-2017
177.80.22.38	Link_visited_7	08:14:25, 18-May-2017

177.80.22.38	Link_visited_4	07:14:16, 19-May-2017

Giai đoạn 2: Xây dựng các chuỗi tuần tự dựa theo các User_IP

Với mỗi User_IP thực hiện các truy cập trong thời gian khác nhau, các chuỗi tuần tự được xây dựng bằng cách biểu diễn các truy cập của từng User_IP (trong ***Bảng 2.2***) theo hàng ngang như sau:

Sequence 1 : Link_visited_1 -1 Link_visited_4 -1 Link_visited_5 -1 -2

Sequence 2: Link_visited_3 -1 Link_visited_2 -1 Link_visited_5 -1 Link_visited_6 -1 -2

Sequence 3: Link_visited_2 -1 Link_visited_3 -1 Link_visited_1 -1 -2

Sequence 4: Link_visited_7 -1 Link_visited_4 -1 -2

Trong đó, các chuỗi tuần tự *Sequence 1, Sequence 2, Sequence 3, Sequence 4* tương ứng với từng User_IP trong cơ sở dữ liệu Web Log trên. Kí hiệu *-1* dùng để phân tách các truy cập Web. Kí hiệu *-2* để biểu diễn sự kết thúc của một chuỗi tuần tự.

Chi tiết giải thuật

Giải thuật biến đổi cơ sở dữ liệu Web Log thành cơ sở dữ liệu tuần tự của luận án được trình bày trong công trình nghiên cứu [CT3]. Chi tiết của giải thuật như sau:

Dữ liệu nhập vào: Một thư mục chứa các tập tin Web log (cơ sở dữ liệu Web Log)

Dữ liệu thu được: Một danh sách các chuỗi dữ liệu tuần tự (một cơ sở dữ liệu tuần tự).

Bước 1: Mở kết nối với cơ sở dữ liệu Web Log

Bước 2: Thực thi vấn tin lấy các thuộc tính User_IP và thuộc tính *Link_visited* từ thư mục chứa các tập tin Web Log.

Bước 3: Thực hiện giải thuật xây dựng cơ sở dữ liệu tuần tự với Mã giả (Pseudo Code) như sau:

Khai báo các biến:

+ *Arr_WebLog* là mảng chứa các mẫu tin của cơ sở dữ liệu WebLog có được bằng cách truy vấn các tập tin Web Log, những mẫu tin trùng lắp, dư thừa bị loại bỏ.

+ *N* là số lượng các mẫu tin chứa trong mảng *Arr_WebLog*.

+ *Arr_User_IP* là mảng một chiều chứa các địa chỉ IP người dùng Web *User_IP*.

+ *Arr_Link* là mảng một chiều chứa các liên kết truy cập.

+ *Arr_Distinct_User_IP* là mảng một chiều lưu các giá trị User_IP khác nhau

+ *Arr_Distinct_Link* là mảng một chiều lưu các giá trị liên kết truy cập khác nhau *Link_visited*

```
1. N ← Length(Arr_WebLog)
2. Arr_User_IP ←  null
3. Arr_Link ← null
4. for i = 0 to N-1 do
5.        Arr_User_IP(i) ← các giá trị của thuộc tính User_IP
6.        Arr_Link(i) ← các giá trị của thuộc tính Link_visited
7. end for
8. Arr_Distinct_User_IP ← arr_Link.Distinct().ToArray();
9. Arr_Distinct_Link ← arr_Link.Distinct().ToArray();
10. count : Số lượng liên kết truy cập của người dùng.
11. count ←  i
12. for k = 0 to count do
13.       for l = 0 to Count(Arr_Distinct_Link) do
14                 if Arr_Link(k) ←  Arr_Distinct_Link(l) then
15.                Arr_Link(k) ←  Arr_Link(k)  + " -1 "
16.           // " –1 " :  Ký hiệu phân cách giữa hai liên kết truy cập liên tiếp
17.       end for
```

18. **end for**
19. **for** *j* = 0 to *count – 1* **do**
20. **if** *Arr_User_IP(j)* < > *Arr_User_IP(j + 1)* **then**
21. *Arr_Link(j)* ← *Arr_Link (j)* + " –2 \r\n"
22. // "-2": Ký hiệu kết thúc một chuỗi tuần tự trong cơ sở dữ liệu tuần tự
23. **end for**
24. *List* ← null: Khởi tạo mảng một chiều để lưu trữ các chuỗi tuần tự
25. **for** *j* = 0 to *count* **do**
26. //Chọn các chuỗi tuần tự có từ 3 liên kết truy cập trở lên:
27. **if** (Number_of_Links ≥ 3)
28. Add *Arr_Link(j)* to *List*
29. **end for**
30. Xuất ra kết quả: Một danh sách các chuỗi tuần tự (Một cơ sở dữ liệu tuần tự)

Giải thích mã giả của giải thuật xây dựng cơ sở dữ liệu tuần tự từ cơ sở dữ liệu Web Log

* Từ Dòng 1 đến Dòng 9: Tìm một mảng chứa các User_IP khác nhau (gọi mảng này là *Arr_Distinct_User_IP* và một mảng một chiều chứa các liên kết truy cập khác nhau (gọi mảng này là Arr_Distinct_Link).

* Từ *Dòng 10* đến *Dòng 18*: Xác định mảng chứa các liên kết truy cập khác nhau, mỗi truy cập được phân cách nhau bằng ký hiệu "*-1"*. Mỗi liên kết truy cập được mã hóa chính bằng số thứ tự của nó trong danh sác các liên kết truy cập khác nhau.

* Từ *Dòng 19* đến *Dòng 23*: Với mỗi địa chỉ IP người dùng khác nhau có một nhóm các truy cập khác nhau. Mỗi nhóm được phân cách nhau bởi ký hiệu *"-2"*. Những nhóm này chính là các chuỗi tuần tự trong cơ sở dữ liệu tuần tự.

* Từ *Dòng 24* đến *Dòng 29*: Các chuỗi tuần tự lần lượt được đưa vào một danh sách để tạo nên cơ sở dữ liệu tuần tự.

Bên cạnh đó, có thể cải tiến giải thuật để tốc độ thực hiện được hiệu quả hơn (với sự hỗ trợ của máy tính có CPU đa nhân, đa luồng xử lý, trong trường hợp này nghiên cứu sinh sử dụng CPU có nhân với 8 luồng xử lý song song), có thể chỉnh sửa giải thuật để thực thi tốt hơn bằng cách điều chỉnh giải thuật từ *dòng 12* đến *dòng 23* như sau:

12. (Parallel) for *k* = 0 to *count* do
13. (Parallel) for l = 0 to Count(Arr_Distinct_Link) do
14 if *Arr_Link(k)* ← *Arr_Distinct_Link(l)* then
15. *Arr_Link(k)* ← *Arr_Link(k)* + " -1 "
16. // " –1 " : Ký hiệu phân cách giữa hai liên kết truy cập liên tiếp
17. end (Parallel) for
18. end (Parallel) for
19. (Parallel) for *j* = 0 to *count – 1* do
20. if *Arr_User_IP(j)* < > *Arr_User_IP(j + 1)* then
21. *Arr_Link(j)* ← *Arr_Link (j)* + " –2 \r\n"
22. // "-2": Ký hiệu kết thúc một chuỗi tuần tự trong cơ sở dữ liệu tuần tự
23. (Parallel) end for

**** Đánh giá độ phức tạp của giải thuật***

- ✓ Các dòng lệnh 1, 2, 3 tốn chi phí thời gian là O(1).

- ✓ Các dòng lệnh 4, 5 tốn chi phí thời gian là O(1).
- ✓ Vòng lặp 4 thực hiện N lần, mỗi lần tốn chi phí thời gian O(1), vậy tổng chi phí thời gian để chương trình thực hiện vòng lặp 4 là O(N).
- ✓ Các dòng lệnh 8, 9, 10, 11 tốn chi phí thời gian là O(1).
- ✓ Dòng lệnh 15 tốn chi phí thời gian là O(1), do đó lệnh điều kiện 14 tốn chi phí thời gian là O(1).
- ✓ Vòng lặp 13 thực hiện *(Count(Arr_Distinct_Link) + 1)* lần, mỗi lần tốn chi phí thời gian là O(1), vòng lặp 12 thực hiện *(count + 1)* lần, do đó tổng chi phí thời gian để chương trình thực hiện vòng lặp 12:
  ((count + 1)* (Count(Arr_Distinct_Link) + 1)).
- ✓ Dòng lệnh 21 tốn chi phí thời gian là O(1) do đó lệnh điều kiện 20 tốn chi phí thời gian là O(1).
- ✓ Dòng lặp 19 thực hiện *count* lần, mỗi lần tốn chi phí thời gian là O(1) nên độ phức tạp để chương trình thực hiện vòng lặp 19 là *O(count).*
- ✓ Dòng lệnh 24 tốn chi phí thời gian là *O(1)*
- ✓ Dòng lệnh 28 tốn chi phí thời gian là *O(1)* do đó dòng lệnh điều kiện 27 tốn chi phí thời gian *là O(1)*

Tổng hợp phân tích bên trên, độ phức tạp của giải thuật là:

O(N) + O((count + 1)* (Count(Arr_Distinct_Link) + 1)).

Độ phức tạp này phụ thuộc vào các tham số *N, count, Count(Arr_Distinct_Link)*, do đó độ phức tạp của giải thuật chuẩn hóa cơ sở dữ liệu tuần tự: *O (Max (N, (count+ 1)* (Count(Arr_Distinct_Link))))*.

Nhận xét rằng Count(Arr_Distinct_Link) ≤ count ≤ N, vì vậy, trong trường hợp xấu nhất độ phức tạp của giải thuật này là *N = Count = Count(Arr_Distinct_Link),* khi đó độ phức tạp của giải thuật là O(Max(N,(N+1)(N)) = O(N^2).

- Trong trường hợp tốt nhất, Count = *Count(Arr_Distinct_Link) = 2,* khi đó độ phức tạp của giải thuật là O(Max(N,2)) = O(N).

2.3.4.4. Các độ đo đánh giá

Độ đo đánh giá về thời gian:

Nếu $t_{non\text{-}parallel}$ lớn hơn rất nhiều so với $t_{parallel}$: Việc xây dựng các cơ sở dữ liệu tuần tự hiệu quả về thời gian, ngược lại thì không hiệu quả về thời gian.

2.3.4.5. Các kết quả thử nghiệm

Giải thuật được thực hiện trên một máy tính cá nhân với cấu hình như sau:

**** Cấu hình phần cứng***

RAM: 32 GB (31.6 GB usable); Intel(R) Core(TM) i7-4800MQ CPU @ 2.70GHz.

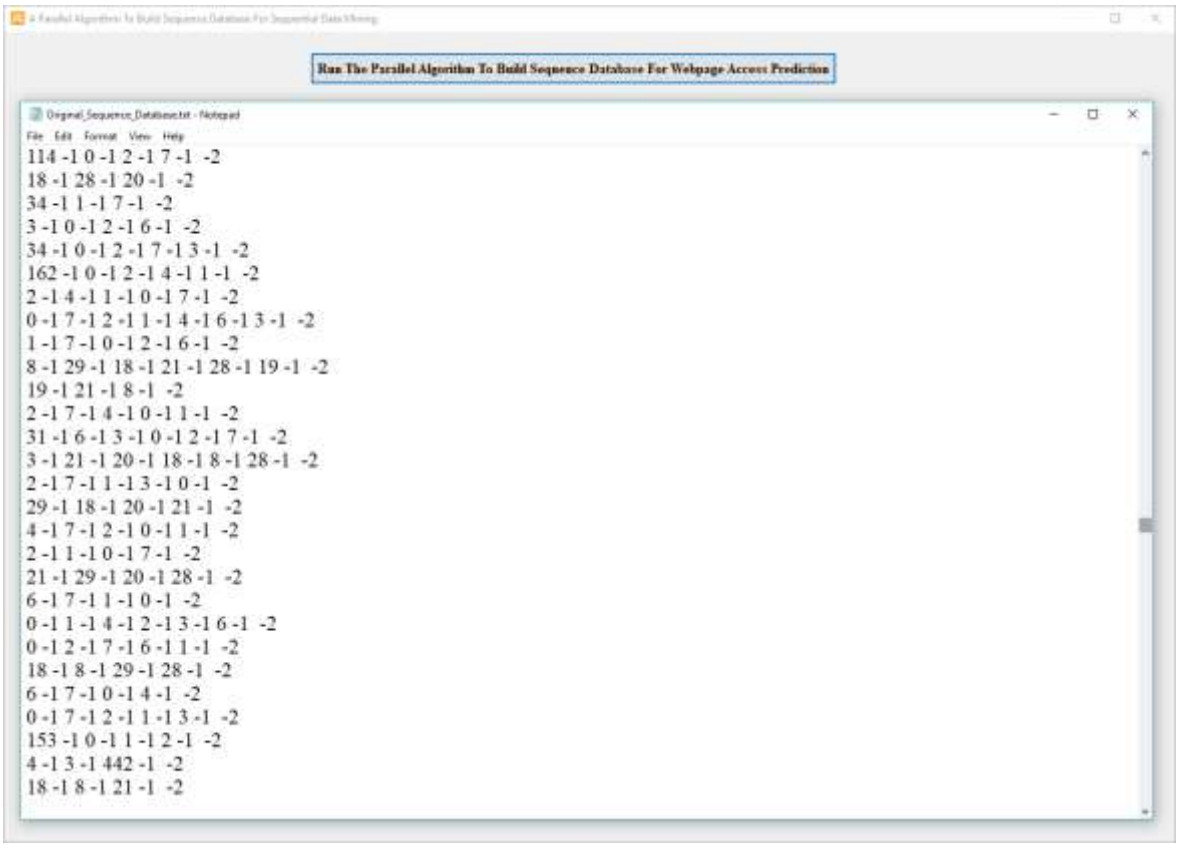
**** Cấu hình phần mềm***

Hệ điều hành 64-bit Windows 10 Education.

Môi trường lập trình C# 2013, thư viện Log Parser Studio 2.2.

**** Các kết quả thử nghiệm:***

Cơ sở dữ liệu tuần tự của dữ liệu nhật ký truy cập Web được minh họa ở ***Hình 2.1***.

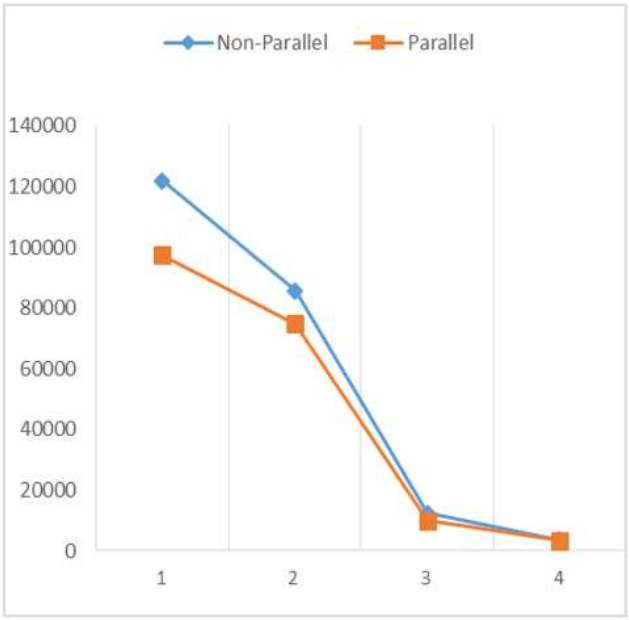


Hình 2.1 *Cơ sở dữ liệu tuần tự của dữ liệu nhật ký truy cập*

Bảng 2.4 trình bày sự so sánh về thời gian thực hiện giải thuật chuẩn hóa cơ sở dữ liệu tuần tự bằng hai phương pháp: xử lý tuần tự và xử lý song song trên Web Log của 4 Website 1, 2, 3, 4 như đã mô tả ở trên.

Bảng 2.4 *So sánh thời gian thực hiện giải thuật xây dựng cơ sở dữ liệu tuần tự*

	Website 1	**Website 2**	**Website 3**	**Website 4**
Non-Parallel (Thời gian thực thi giải thuật tuần tự) *(milliseconds)*	121836	85683	12382	3508
Parallel (Thời gian thực thi giải thuật song song) *(milliseconds)*	97449	74814	9893	3312
Số lượng chuỗi tuần tự được tạo	12211	9678	312	330



Hình 3.2 *So sánh thời gian thực thi giải thuật tuần tự và song song*

Hình 2.2 cho thấy khi kích cỡ của cơ sở dữ liệu Web Log càng lớn thì khoảng cách về thời gian thực thi bằng tiếp cận tuần tự và song song của giải thuật xây dựng cơ sở dữ liệu tuần tự càng cao. Điều đó có nghĩa là với cơ sở dữ liệu Web Log càng lớn xử lý tính toán song song cho giải thuật sẽ cho hiệu quả tối ưu hơn. Nghiên cứu sinh cũng đã công bố một số công trình liên quan đến nghiên cứu này là công trình [CT2], [CT3] và [CT6]. Ngoài ra, nghiên cứu liên quan đến thiết kế cơ sở dữ liệu tuần tự từ cơ sở dữ liệu có nhãn thời gian (temporal networks) cũng đã được nghiên cứu sinh thực hiện trong công trình nghiên cứu [CT8].

2.3.5. Đánh giá và thảo luận

Các kết quả thực nghiệm trên đã trình bày cách thức xây dựng các cơ sở dữ liệu tuần tự để dự đoán truy cập Web bằng hai phương pháp xử lý tuần tự và song song.

Bên cạnh đó, một vấn đề được đặt ra là việc xây dựng và chuẩn hóa các cơ sở dữ liệu có thực sự cần thiết? Để tìm câu trả lời cho câu hỏi này, số liệu trong *Bảng 2.5* cho thấy rằng có sự chênh lệch rất lớn về số lượng các mẫu tin trong các cơ sở dữ liệu Web Log so với số lượng các mẫu tin trong các cơ sở dữ liệu tuần tự trên 4 Website được nghiên cứu.

Bảng 2.5 Độ tương quan về số lượng mẫu tin giữa cơ sở dữ liệu Web Log và cơ sở dữ liệu tuần tự

	Số mẫu tin cơ sở dữ liệu Web Log	**Số mẫu tin cơ sở dữ liệu tuần tự**
Website 1 *periwinklelecottages.com* [1]	3511237	12211

[1] Truy cập ngày 22/8/2017

Website 2 *palmviewsanibel.com* [1]	4217568	9678
Website 3 *devqa.robotec.co.il* [2]	2527429	312
Website 4 *inees.org* [3]	593367	330

Cụ thể, trong Website thứ nhất, cơ sở dữ liệu tuần tự thu được chỉ có 12211 mẫu tin, chỉ xấp xỉ 1/287 so với số lượng mẫu tin trong cơ sở dữ liệu Web Log của cùng Website.

Tương tự, trong Website thứ hai, cơ sở dữ liệu tuần tự thu được chỉ có 9678 mẫu tin, chỉ xấp xỉ 1/435 so với số lượng mẫu tin trong cơ sở dữ liệu Web Log của cùng Website.

Hai ví dụ còn lại ở Website thứ ba và Website thứ tư, các cơ sở dữ liệu tuần tự thu được có số mẫu tin là không đáng kể so với cơ sở dữ liệu Web Log của các Website này.

Số lượng các mẫu tin thu được trong các cơ sở dữ liệu tuần tự là không đáng kể so với số mẫu tin trong các cơ sở dữ liệu Web Log. Điều này cho thấy cơ sở dữ liệu tuần tự đã được loại bỏ những thông tin dư thừa không cần thiết. Như vậy, cơ sở dữ liệu thu được từ cơ sở dữ liệu Web Log đem lại nhiều lợi ích: (1) Không gian dự đoán được thu hẹp giúp cho thời gian thực hiện dự đoán truy cập Web được tốt hơn.

[1] Truy cập ngày 22/8/2017

[2] Truy cập ngày 23/8/2017

[3] Truy cập ngày 23/8/2017

(2) Việc dự đoán sẽ chính xác hơn khi những dữ liệu dư thừa, không phục vụ cho dự đoán được loại bỏ trước khi áp dụng các giải pháp dự đoán truy cập Web.

2.3.6. Kết luận chương 2

Cơ sở dữ liệu tuần tự cho dự đoán truy cập Web được thu thập từ dữ liệu click-stream hoặc được chuẩn hóa và xây dựng từ nhật ký Web. Trong chương này, luận án đã trình bày các tiếp cận để xây dựng cơ sở dữ liệu tuần tự phục vụ cho dự đoán truy cập Web. Cụ thể, nghiên cứu sinh đã đề xuất một giải pháp khác nhau để thiết kế cơ sở dữ liệu tuần tự từ cơ sở dữ liệu nhật ký Web. Ngoài ra, nghiên cứu sinh cũng thực hiện các công trình nghiên cứu liên quan về chủ đề này như thiết kế cơ sở dữ liệu tuần tự cho mạng có nhãn thời gian [CT8].

CHƯƠNG 3. NÂNG CAO HIỆU QUẢ VỀ ĐỘ CHÍNH XÁC KHAI PHÁ DỮ LIỆU TUẦN TỰ CHO DỰ ĐOÁN TRUY CẬP WEB

3.1. Giới thiệu

Chương 3 trình bày một giải pháp tích hợp giải thuật PageRank với CPT+ để nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Dữ liệu đầu vào cho nghiên cứu là các cơ sở dữ liệu tuần tự được thu thập từ các tập dữ liệu thu thập từ các tập dữ liệu click-stream, cụ thể là các cơ sở dữ liệu tuần tự FIFA, KOSARAK, MSNBC[1]. Tuy nhiên, những cơ sở dữ liệu tuần tự này cần được cải thiện thêm về độ chính xác vì các cơ sở dữ liệu tuần tự này còn ẩn chứa nhiều dữ liệu dư thừa và không có ý nghĩa cho dự đoán truy cập Web. Bằng giải pháp áp dụng kỹ thuật tính toán PageRank cho các chuỗi dữ liệu tuần tự kết hợp với CPT+, nghiên cứu sinh thu được các cơ sở dữ liệu tuần tự có độ chính xác cao hơn để hỗ trợ cho dự đoán truy cập Web tốt hơn về độ chính xác.

3.2. Cơ sở lý luận của giải pháp

Một số lý do tính toán PageRank được chọn cùng với CPT+ để nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web:

(1) Thuật toán PageRank là một thuật toán nổi tiếng và có nhiều ứng dụng:

- ✓ Trong lãnh vực thể thao, giải thuật PageRank được dùng để xếp hạng thành tích thi đấu cho các đội bóng trong giải nhà nghề Hoa Kỳ (NFL: National Football League) [116], xếp hạng các vận động viên thể thao ở giải điền kinh thế giới (Diamond League) [11].
- ✓ PageRank được dùng trong mạng xã hội Twitter để xếp hạng những người dùng theo dõi các bài viết[WTF: The Who to Follow Service at Twitter].

[1] https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

- ✓ Một phiên bản gần đây của PageRank được ứng dụng để tính chỉ số trích dẫn và tầm quan trọng của các bài báo khoa học ISI và Scopus [14].

(2) Theo nghiên cứu [73], dựa trên giả định là các liên kết truyền đạt các khuyến nghị của con người có thể được rút ra trực tiếp, nhiều người đã tiến hành nghiên cứu về phân tích liên kết (Chẳng hạn PageRank [85] và HITS [65]) để khai thác cấu trúc Web nhằm nắm bắt tầm quan trọng của một trang Web. Cũng theo nghiên cứu này, so với các kỹ thuật khai phá dữ liệu văn bản[Modern information retrieval], chỉ xếp hạng phụ thuộc truy vấn, các cách tiếp cận xem xét phân tích liên kết có thể cung cấp xếp hạng phụ thuộc truy vấn đề tìm kiếm trang web chất lượng cao trong các công cụ tìm kiếm web toàn cầu [5; 85].

(3) Về bản chất, PageRank diễn giải một siêu liên kết từ trang pageA đến trang pageB dưới dạng phiếu bầu [112]. Bên cạnh đó, PageRank cũng phân tích các trang được bỏ phiếu nghĩa là các trang nhận được nhiều liên kết đến (tương tự như nhận được nhiều phiếu bầu) thì càng có ý nghĩa quan trọng hơn và có độ tin cậy cao hơn những trang có ít liên kết đến hơn. Như vậy mối quan hệ giữa các trang với nhau có tầm quan trọng đặc biệt. Bên cạnh đó, xét các chuỗi dữ liệu tuần tự, một mối quan hệ tương tự mối quan hệ trên giữa các trang với nhau chính là thứ tự các trang được truy cập trước và sau theo thứ tự thời gian. Với nhận định này, việc thử nghiệm tính toán PageRank kết hợp với CPT+ cho dự đoán truy cập Web theo trực giác có thể sẽ mang đến những kết quả khả quan và các kết quả thực nghiệm sẽ được trình bày ở các phần sau đã chứng minh giải pháp dự đoán truy cập Web dựa trên giải thuật PageRank và CPT+ là một giải pháp khả thi.

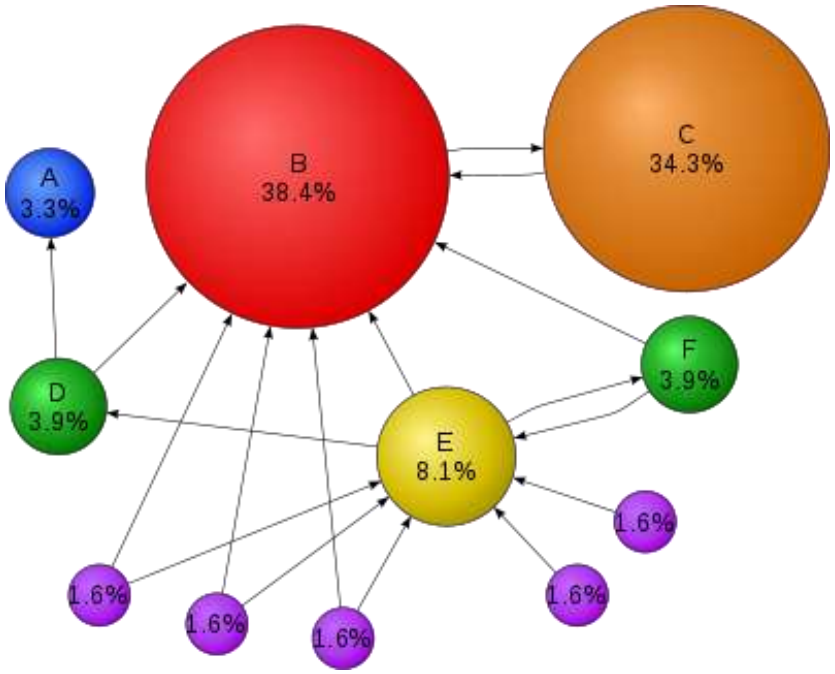
Ý tưởng của giải pháp nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web được đề xuất là khai thác tầm quan trọng của các liên kết được truy cập bởi người dùng Web (dựa trên thế mạnh của giải thuật PageRank) và tận dụng sức mạnh của dự đoán chuỗi dữ liệu tuần tự (dựa trên ưu điểm của giải thuật CPT+) để làm giảm thời gian dự đoán truy cập Web. Nội dung

của giải pháp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web sẽ được trình bày trong phần tiếp theo.

Kết quả hướng đến của giải pháp là sắp xếp để chia các chuỗi tuần tự thành hai phần: Phần trên là tập hợp các chuỗi tuần tự có trung bình chỉ số PageRank cao, phần dưới là tập hợp các chuỗi tuần tự có trung bình chỉ số PageRank thấp. Một điều hiển nhiên là khi một trang Px có chỉ số PageRank cao thì số trang dự đoán ra trang Px càng cao (vì khi Px có PageRank cao thì số liên kết trỏ đến nó càng cao). Như vậy, đây chính là yếu tố góp phần cho việc dự đoán thành công khi tích hợp giải pháp với CPT+.

3.3. Nội dung của giải pháp nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web

Tính toán PageRank dựa trên ý tưởng đếm backlinks (trích dẫn) đến một trang nhất định [17]. Nó được đề xuất bởi Sergey Brin và Lawrence Page và thuật toán này cung cấp một phương pháp để đo tầm quan trọng của các trang web. ***Hình 3.1*** minh họa cách tính PageRank.



Hình 3.1 *Một ví dụ trực quan về PageRank* [1]

Giải thuật PageRank là cơ sở để xây dựng công cụ tìm kiếm nổi tiếng và thành công là Google [112]. Theo nghiên cứu [17], PageRank có thể được tính toán dùng một giải thuật lặp đơn giản, và nó liên quan đến vector riêng của ma trận liên kết chuẩn hóa của Web. Hơn nữa, nghiên cứu [112] chỉ ra rằng giải thuật PageRank là 1 trong 10 giải thuật quan trọng nhất của khai phá dữ liệu [112]. Các nhà phát triển giải thuật PageRank đưa ra một công thức để tính chỉ số PageRank của một trang A (liên kết A) như sau:

$$PR(page\ A) = (1-df) + df(PR(T_1)/C(T_1) + PR(T_2)/C(T_2) + ... + PR(T_n)/C(T_n)) \quad (3.1)$$

Trong đó

PR(page A): Chỉ số PageRank của trang Web A

T_i*:* Một trang liên kết đến trang A

$PR(T_i)$*:* Chỉ số PageRank của trang T_i

$C(T_i)$*:* Số lượng các trang mà T_i liên kết đến

df: Chỉ số damping factor (*df* = 0.85 được nhiều nhà nghiên cứu sử dụng): Hệ số điều chỉnh. Tham số này cho biết xác suất của một người dùng ngẫu nhiên liên tục truy cập vào liên kết trên trang khi họ truy cập vào một Website.

3.4. Giải pháp nâng cao độ chính xác dự đoán truy cập Web với giải thuật PageRank và CPT+

Mô tả giải thuật:

Dữ liệu nhập vào:

Cơ sở dữ liệu tuần tự

Dữ liệu thu được:

Cơ sở dữ liệu tuần tự thu gọn.

---

(1) https://en.wikipedia.org/wiki/PageRank

- Thủ tục ***Build_GraphDatabase***

Đây là thủ tục biến đổi mô hình cơ sở dữ liệu tuần tự sang mô hình cơ sở dữ liệu đồ thị (Graph Database). Trong đó mô hình Graph Database là mô hình cơ sở dữ liệu đồ thị với mỗi liên kết trong cơ sở dữ liệu tuần tự là một nút (node) và liên kết ngay sau liên kết đó là một nút kề với nút đó. Giữa hai nút với nhau được biểu diễn bằng một đường nối như minh họa ở Hình 3.2.

Gọi *arr* là mảng với các phần tử là các chuỗi dữ liệu tuần tự trong cơ sở dữ liệu tuần tự được tạo từ thủ tục ***Clean_SequenceDatabase***

sfile là chuỗi lưu các hàng của ma trận kề của của cơ sở dữ liệu đồ thị.

n1 là mảng một chiều lưu các giá trị khác nhau của cơ sở dữ liệu tuần tự *SD*.

Chi tiết thủ tục ***Build_GraphDatabase*** được minh họa như sau:

```
Procedure Build_GraphDatabase
Begin
Input: Cơ sở dữ liệu tuần tự
1. String sfile ← null;// Khởi tạo chuỗi sfile là chuỗi rỗng
2. Sort(n1); //Sắp xếp tăng dần các phần tử trong mảng n1
3.    For k ← 0 to Len(n1) - 1 do
4.    Begin
5.                  sfile ← sfile + n1[k] +" ";
6.          For i ← 0  to  Len(arr) -1 do
7.                Begin
8.                  For j←0 to j < Len (arr[i]) -1 do
9.                    If (arr[i][j] = n1[k]) Then
10.                             sfile ←sfile+arr[i][j+1]+ " ";
11.               End
12.             sfile  ←  sfile +"\n";
13.    End
```

14. **WriteFile** sfile Adjacency_Matrix
Output: Ma trận kề các nút trong cơ sở dữ liệu đồ thị
End

Giải thích thủ tục ***Build_GraphDatabase***:
+ Vòng lặp từ *Dòng 300* đến Dòng *13*: Duyệt mảng n1
+ *Dòng 5:* Thêm phần tử vào chuỗi kết quả
+ Vòng lặp từ *Dòng 6* đến *Dòng 11:* Duyệt mảng arr
+ *Dòng 8* đến *Dòng 10*: Xác định mỗi đỉnh kề với đỉnh nào
+ *Dòng 14*: Ghi tập tin kết quả vào file *Adjacency_Matrix*

Kết quả của thủ tục là một tập tin chứa ma trận kề chứa các mối quan hệ kề nhau giữa các đỉnh trong cơ sở dữ liệu đồ thị.

Tính toán giá trị PageRank từng đỉnh cho cơ sở dữ liệu đồ thị.

Dữ liệu nhập vào: Đồ thị biểu diễn các liên kết (Cơ sở dữ liệu đồ thị)

Dữ liệu thu được: Mảng n phần tử lưu các giá trị PageRank của mỗi liên kết

localPR: mảng lưu giá trị tăng lên của giá trị pagerank bên trong mỗi chunk

danglingContrib: biến lưu giá trị đóng góp của đỉnh dangling (đỉnh không liên kết với bất kỳ đỉnh nào)

globalPR: Tổng hợp các giá trị *localPR*

tempRecv: Tổng hợp giá trị *danglingContrib*

df: Hằng số Damping Factor có giá trị thường là 0.85 (Damping Factor được xem xét khi có người dùng click ngẫu nhiên vào các liên kết và cuối cùng dừng lại. Xác suất mà một người dùng sẽ tiếp tục là một Damping Factor và thường được gán giá trị là 0.85).

Giải thích giải thuật tính ***PageRank*** (xem mã nguồn ở Phụ lục 3)

Mã giả của thuật toán này được nghiên cứu sinh xây dựng theo nghiên cứu của Mridul Birla, Kai Zhen [(1)].

Gọi adjMatrix là bản băm để lưu trữ các đỉnh kề nhau.

Mỗi đỉnh của đồ thị sẽ được tính theo công thức sau:

$$PR(u) = \frac{1-d}{n} + d * \sum v \in Set \frac{PR(v)}{L(v)} \quad (3.3)$$

Vòng lặp từ dòng lệnh *1* đến dòng lệnh *65* để duyệt chỉ số lặp, giả sử *iterations* = 10 thì phép lặp thực hiện 10 lần.

* Dòng *3* đến dòng *4*:

Khởi tạo *localPR* của mỗi đỉnh đồ thị và khởi tạo *danglingContrib*

* Vòng lặp *6* đến *21*: Duyệt bản đồ băm ma trận kề

* Dòng lệnh *8* đến dòng lệnh *10*: Tính toán *danglingContrib*

* Dòng lệnh *11* đến dòng lệnh *20*: Tính toán giá trị PageRank mỗi đỉnh

* Dòng lệnh *22* đến dòng lệnh *24*: Khởi tạo các thông số gửi và nhận thông tin cho tính toán song song PageRank

* Dòng lệnh *25*: Gửi thông tin danglingContrib cho các máy khách

* Dòng lệnh *26*: Gửi thông tin PageRank cho các máy khách

* Dòng lệnh *27* đến dòng lệnh *35*: Sau khi gửi thông tin để các máy khách tính toán, máy chủ sẽ tổng hợp kết quả.

- Thủ tục ***Average_by_sequences***

Thủ tục ***Average_by_sequences*** sẽ xác định giá trị trung bình của các chuỗi tuần tự trong cơ sở dữ liệu tuần tự.

Đặt *arr_avg* là mảng chứa các giá trị trung bình PageRank của từng chuỗi tuần tự chứa các liên kết có trong cơ sở dữ liệu tuần tự.

[1] Birla, Kai Zhen (Distributed System, github.com/cocosci/MPIPagerank, truy cập ngày 17/06/2018)

Đặt *arr_temp* là mảng chứa các giá trị PageRank của từng chuỗi tuần tự chứa các liên kết có trong cơ sở dữ liệu tuần tự. Chi tiết của thủ tục ***Average_by_sequences*** được trình bày bằng mã giả như sau:

```
Procedure Average_by_sequences
Begin
1. For i ← 0 to Len(arr_temp) - 1 do
2. arr_avg[i] ← Average_Rows (arr_temp, Len(arr_temp))
End
```

Trong đó hàm ***Average_Rows*** được cài đặt như sau:

```
Function Double Average_Rows(Double arr[][],int n, int k)
Begin
1. Double S ← 0.0;
2. Double average ← 0.0;
3. For j ← 0 to Len(arr[k]) - 1 do
4.     Begin
5.         S ← S +  arr[k][j];
6.         average ← S / Len(arr[k]);
7.     End
 Return average;
 End
```

Giải thích thủ tục hàm ***Average_Rows***:

Dòng lệnh 1, Dòng lệnh 2: Khởi tạo S chứa giá trị tổng các phần tử trên hàng.

Dòng lệnh 3: Vòng lặp duyệt cơ sở dữ liệu tuần tự (có dạng mảng răng cưa - jagged array.

Dòng lệnh 5: Tính các tổng các giá trị trên từng chuỗi tuần tự.

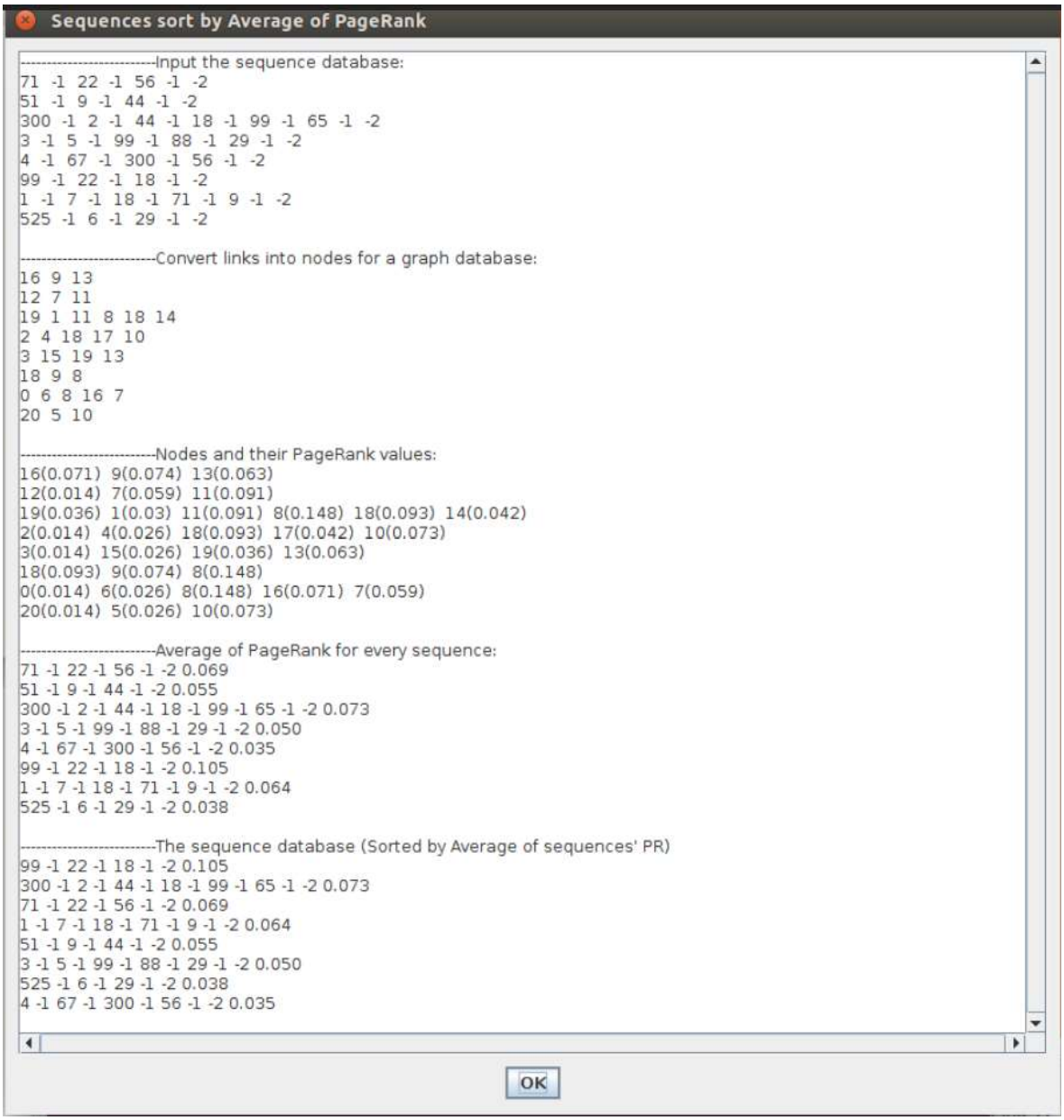
Dòng lệnh 6: Xác định trung bình các giá trị trên từng chuỗi tuần tự.

- Thủ tục ***Sort_Sequences***

Thủ tục ***Sort_Sequences***: Sắp xếp các chuỗi tuần tự theo giá trị trung bình PageRank của mỗi chuỗi tuần tự từ cao xuống thấp.

Gọi arr là mảng răng cưa (jagged array) chứa các chuỗi tuần tự trong cơ sở dữ liệu tuần tự, các phần tử của mảng này chính là các liên kết mà người dùng truy cập.

```
Procedure Sort_Sequences
Begin
1. For i ← 1 to Len(arr) -1 do
2. For (j ← Len(arr) – 1 to i; j ← j - 1)
3.     If  arr_avg[i] > arr_avg[j] Then
4.         Begin
5.             temp = arr[j];
6.           arr[j] = arr[j-1];
7.           arr[j-1]=temp;
8.         End
End
```



Hình 3.2 *Tính toán từng bước giá trị trung bình PageRank của các chuỗi tuần tự*

*** Đánh giá độ phức tạp của giải thuật**

Độ phức tạp của giải thuật được tính toán dựa theo các chi phí thời gian thực hiện các mã nguồn *[Source 1] , [Source 2] , [Source 3] , [Source 4] , [Source 5]* ở phần PHỤ LỤC 1.

Gọi *L* là chiều dài trung bình của chuỗi tuần tự, hay còn gọi là số lượng trung bình của các chuỗi tuần tự, K là số các phần tử của cơ sở dữ liệu tuần tự (các phần tử có thể trùng nhau), K' là số lượng các phần từ đôi một khác nhau trong cơ sở dữ liệu tuần tự, *arr* là mảng răng cưa (jagged array) với các phần tử là các chuỗi tuần tự, *arr_temp* là mảng tạm của mảng răng cưa *arr*.

- Xét [Source 1] (PHỤ LỤC 1):
  - ✓ Dòng lệnh 1 tốn chi phí thời gian là O(1).
  - ✓ Dòng lệnh 5 tốn chi phí thời gian là O(1)
  - ✓ Vòng lặp 4 thực hiện *L* lần, mỗi lần tốn chi phí thời gian là O(1) và vòng lặp 2 thực hiện *N* lần (*N* là số chuỗi tuần tự trong cơ sở dữ liệu tuần tự). Do đó độ phức tạp để chương trình thực hiện vòng lặp 2 là *O(N*L).*
  - ✓ Dòng lệnh 7, 8, 11 tốn chi phí thời gian là O(1)
  - ✓ Vòng lặp 9 thực hiện K lần. Do đó độ phức tạp để chương trình thực hiện vòng lặp 9 là O(K).
  - ✓ Dòng lệnh 13 tốn chi phí thời gian là O(K)
  - ✓ Dòng lệnh 14 tốn chi phí thời gian là O(K') (Với)

Vậy độ phức tạp để thực hiện *[Source 1]* là O(Max(O(N*L), O(K),O(K'). Vì K'<<K, độ phức tạp của *[Source 1]* là O(N*L). Trên thực tế $L \leq N$, do đó, trong trường hợp xấu nhất, L = N, khi đó độ phức tạp của *[Source 1]* là $O(N^2)$,
và trong trường hợp tốt nhất, L = 2 và $L \ll N$, độ phức tạp của *[Source 1]* là O(N).

- Xét [Source 2] (PHỤ LỤC 1):
  - ✓ Dòng lệnh 1 tốn chi phí thời gian là O (1)
  - ✓ Dòng lệnh 2 tốn chi phí thời gian là O(K')

- ✓ Dòng lệnh 4 tốn chi phí thời gian là O(1)
- ✓ Dòng lệnh 10 tốn chi phí thời gian là O(1), do đó lệnh điều kiện 8 tốn chi phí thời gian là O(1)
- ✓ Vòng lặp 7 thực hiện L lần, mỗi lần tốn chi phí O(1), vòng lặp 5 thực hiện N lần và vòng lặp 3 thực hiện *(L-1)* lần, do đó độ phức tạp để chương trình thực hiện vòng lặp 3 là *O(K'*N*(L-1))*

Như vậy độ phức tạp của *[Source 2]* là *O(Max(O(K'), O(K'*N*(L-1)))*. Nói cách khác độ phức tạp của *[Source 2]* là *O(K'*N*(L-1))*.

Trong trường hợp xấu nhất K' = L = N, độ phức tạp của *[Source 2]* là $O(N^3)$

Trong trường hợp tốt nhất L= K'=2 và L ≪ N và K' L ≪ N, độ phức tạp của *[Source 2]* là O(N).

- Xét [Source 3] (PHỤ LỤC 1)
  - ✓ Dòng lệnh 6 và dòng lệnh 7 tốn chi phí thời gian là O(1), do đó lệnh điều kiện 5 tốn chi phí là O(1).
  - ✓ Dòng lệnh 11 tốn chi phí thời gian là O(1).
  - ✓ Vòng lặp 4 thực hiện N lần, mỗi lần tốn chi phí thời gian là O(1), vòng lặp 3 thực hiện L lần và vòng lặp 1 thực hiện N lần. Do đó độ phức tạp để chương trình thực hiện *[Source 3]* là O(N*L*N).

Trong trường hợp xấu nhất L = N, độ phức tạp để chương trình thực hiện *[Source 3]* là $O(N^3)$.

Trong trường hợp tốt nhất L= 2 và L ≪ N, độ phức tạp để chương trình thực hiện *[Source 3]* là $O(N^2)$.

- Xét [Source 4] (PHỤ LỤC 1)
  - ✓ Dòng lệnh 3 tốn chi phí thời gian là O(1)
  - ✓ Vòng lặp 1 thực hiện N lần, mỗi lần tốn chi phí là O(1). Do đó độ phức tạp của *[Source 4]* là O(N)
- Xét [Source 5] (PHỤ LỤC 1)

- ✓ Dòng lệnh 1 có chi phí thời gian là O(1).
- ✓ Các dòng lệnh 5, 6, 7 có chi phí thời gian là O(1). Do đó lệnh điều kiện 4 có chi phí thời gian là O(1).
- ✓ Vòng lặp 3 thực hiện N lần, mỗi lần tốn chi phí thời gian là O(1) và vòng lặp 2 thực hiện N lần do đó độ phức tạp để chương trình thực hiện vòng lặp 2 là O(N*N) = $O(N^2)$.

Như vậy độ phức tạp của *[Source 5]* là $O(N^2)$

Tóm lại độ phức tạp của giải thuật nâng cao độ chính xác cho dự đoán truy cập Web:

- Trong trường hợp xấu nhất, độ phức tạp của giải thuật:

$O(Max(O(N^2), O(N^3), O(N^2), O(N), O(N^2)) = O(N^3)$

- Trong trường hợp tốt nhất, độ phức tạp của giải thuật:

$O(Max(O(N), O(N), O(N^2), O(N), O(N)) = O(N^2)$

3.5. Các kết quả thử nghiệm nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web

3.5.1. Mục tiêu

Giải pháp này nhằm làm giảm kích cỡ của cơ sở dữ liệu tuần tự mà không làm mất tính chính xác, những dữ liệu thừa, thông tin nhiễu sẽ bị loại bỏ, kết quả sẽ thu được cơ sở dữ liệu tuần tự thu gọn hơn. Điều này sẽ giúp làm tăng hiệu quả về thời gian cho dự đoán truy cập Web..

3.5.2. Dữ liệu

Nghiên cứu sinh đã chọn 3 tập dữ liệu được thu thập trực tuyến tại liên kết philippe-fournier-viger.com/spmf/index.php?link=datasets.php (trang Web của GS. Philippe Fournier Viger). Các tập dữ liệu này chứa các chuỗi tuần tự được truy cập bởi người dùng và được mã hóa thành số để thuận tiện cho mục đích khai phá dữ liệu, đặc biệt là dùng cho dự đoán truy cập Web. Chi tiết các tập dữ liệu như sau:

MSNBC là một dữ liệu các truy cập Web (clickstream). Dữ liệu gốc của tập dữ liệu này lên đến 989,818 chuỗi dữ liệu tuần tự được thu thập từ kho dữ liệu UCI (http://archive.ics.uci.edu/ml/index.php). Trong phạm vi nghiên cứu này, nghiên cứu sinh đã sử dụng tập con của tập dữ liệu gốc. Tập dữ liệu MSNBC ở phạm vi nhỏ hơn này chứa 31,790 chuỗi dữ liệu tuần tự, trong đó có 18 trang Web (liên kết khác nhau) đã được mã hóa thành số để thuận tiện cho khai phá dữ liệu.

FIFA là một tập dữ liệu được tạo ra từ các Web Log (nhật ký Web) sự kiện World Cup. Trong nghiên cứu này, nghiên cứu sinh sử dụng một tập dữ liệu có chứa 20,450 chuỗi tuần tự dữ liệu truy cập Web từ Website FIFA World Cup 1998 (http://hita.ee.lbl.gov/html/contrib/WorldCup.html). Tập dữ liệu này có chứa 20,450 chuỗi dữ liệu tuần tự và 2991 trang Web khác nhau.

KOSARAK là một tập dữ liệu các truy cập Web được thu thập từ một cổng thông tin của từ Hungary (http://fimi.ua.ac.be/data). Tập dữ liệu này chứa 69999 chuỗi dữ liệu tuần tự và 19986 trang Web khác nhau. Đây là tập dữ liệu lớn nhất trong thử nghiệm của nghiên cứu sinh.

*** Môi trường thực hiện**

Các thử nghiệm được thực hiện trên một máy vi tính cá nhân có bộ xử lý Intel i7 third-generation. Bộ nhớ RAM: 32 GB, Hệ điều hành Ubuntu 16.04.5 LTS (Xenial Xerus) trên môi trường Java 8.1 kết hợp với Eclipse Neon.3.

3.5.3. Phương pháp

Thực hiện việc biến đổi cơ sở dữ liệu tuần tự sang cơ sở dữ liệu đồ thị (mỗi nút (node) của cơ sở dữ liệu đồ thị được biểu diễn từ các liên kết trong các chuỗi dữ liệu tuần tự trong cơ sở dữ liệu tuần tự) nhằm mục đích tối ưu không gian dự đoán. Sau đó, thực hiện bài toán ngược là biến đổi cơ sở dữ liệu đồ thị trở lại thành cơ sở dữ liệu tuần tự.

Trong phần nghiên cứu này, một giải pháp giảm kích cỡ không gian dự đoán được đề xuất mà vẫn đảm bảo độ chính xác dự đoán truy cập Web. Khi dự đoán

truy cập Web trên tập dữ liệu được thu gọn, thời gian thực thi dự đoán sẽ được cải thiện đáng kể. Ý tưởng của giải pháp là tận dụng thế mạnh của dự đoán đối tượng kế tiếp của phương pháp CPT+ trong dự đoán chuỗi tuần tự và tầm quan trọng cũng như hữu ích của chỉ số PageRank đối với các liên kết trong một Website.

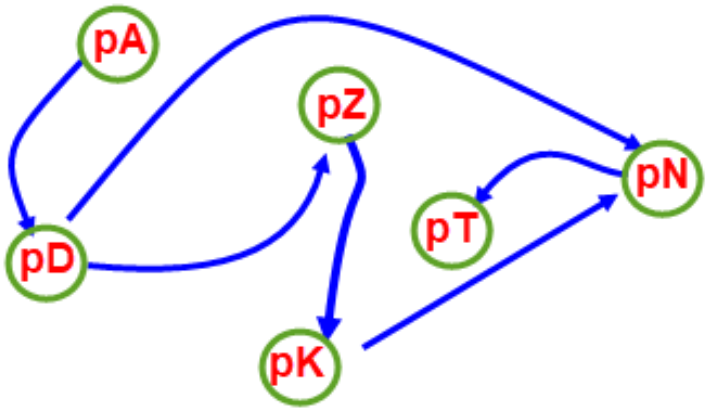
Tham số đầu vào của giải pháp này: Một cơ sở dữ liệu tuần tự (được xây dựng từ dữ liệu Clickstream của bất kỳ Website nào thuộc các chủ đề phân loại Website nào), chỉ số *Damping factor = 0.85* và tỷ lệ phần trăm dữ liệu không bị loại bỏ.

Các bước thực hiện cụ thể như sau:

Bước 1: Biến đổi cơ sở dữ liệu tuần tự thành cơ sở dữ liệu đồ thị.

Giả sử một cơ sở dữ liệu tuần tự *SD* có *N* chuỗi tuần tự. Mỗi cặp liên kết liên tiếp $\{p_i, p_j\}$ theo (trình tự thời gian) trong *SD* có thể được xem như là một mối quan hệ giữa hai đỉnh (nút) của đồ thị có hướng. Trong đó, đường nối p_ip_j xuất phát từ p_i và kết thúc ở p_j là cạnh nối của hai đỉnh này.

Chẳng hạn, giả sử có một cơ sở dữ liệu với hai chuỗi tuần tự sau $S_1 = \langle pA, pD, pZ, pK, pN \rangle$ và $S_2 = \langle pD, pN, pT \rangle$. Đồ thị có hướng biểu diễn cho cơ sở dữ liệu tuần tự này có thể được mô tả như minh họa ở ***Hình 4.2***.



Hình 3.3 *Một đồ thị có hướng được xây dựng từ một cơ sở dữ liệu tuần tự*

Bước 2: Xác định chỉ số PageRank của từng trang

Dựa vào giải thuật PageRank đã được trình bày ở trên, mỗi liên kết trong cơ sở dữ liệu tuần tự sẽ có một chỉ số PageRank tương ứng.

Bước 3: Xác định giá trị trung bình của chỉ số PageRank cho mỗi chuỗi tuần tự

Giả sử rằng cơ sở dữ liệu tuần tự *SD* chứa *N* chuỗi tuần tự, và S_j là chuỗi tuần tự ở vị trí thứ *j* trong *SD*.

Trong cơ sở dữ liệu tuần tự *SD*, với mỗi chuỗi tuần tự, một liên kết trong chuỗi dữ liệu tuần tự có một chỉ số PageRank riêng đã được xác định ở *Bước 2*. Đặt *M* là số liên kết trong chuỗi tuần tự S và p_i là liên kết ở vị trí *i* trong chuỗi tuần tự *S*. Giá trị trung bình các chỉ số PageRank của chuỗi tuần tự *S* được xác định theo công thức sau:

$$AVG\_PR\ (S_j) = \frac{\sum_{M}^{i=1} PR(p_i)}{M} \qquad (3.2)$$

Trong đó:

$AVG\_PR(S_j)$ là giá trị trung bình của tất cả các liên kết có trong chuỗi tuần tự S_j

Bước 4: Sắp xếp tất cả các chuỗi tuần tự trong cơ sở dữ liệu tuần tự SD theo giá trị trung bình của mỗi chuỗi tuần tự từ cao xuống thấp mà vẫn bảo đảm độ chính xác.

Mục đích chính của bước này là loại bỏ các chuỗi tuần tự dư thừa ra khỏi cơ sở dữ liệu tuần tự và chỉ giữ lại các chuỗi tuần tự có ích phục vụ cho dự đoán truy cập Web.

Đặt $k \in (0,100)$ là tỷ lệ phần trăm của kích cỡ cơ sở dữ liệu tuần tự. Chẳng hạn, với $k = 75$ (%) và kích cỡ cơ sở dữ liệu tuần tự là $N = 100000$ thì kích cỡ cơ sở dữ liệu mới sau khi được thu gọn là *75000*.

Để giảm kích cỡ của cơ sở dữ liệu tuần tự, k có thể được chọn một cách ngẫu nhiêu. Tuy nhiên, để bảo toàn độ chính xác của dự đoán chuỗi tuần tự, các giá

trị k thích hợp được chọn. Cụ thể là, đặt acc_1 là độ chính xác của dự đoán chuỗi tuần tự cho cơ sở dữ liệu tuần tự gốc. Tương tự, đặt acc_2 là độ chính xác của dự đoán chuỗi dữ liệu tuần tự của cơ sở dữ liệu được thu gọn. Nếu $acc_2 \geq acc_1$, giá trị k được chọn là hữu dụng. Như vậy k (%) các chuỗi tuần tự trong cơ sở dữ liệu gốc được giữ lại.

Bước 5: Áp dụng mô hình CPT+ để dự đoán chuỗi tuần tự

Với cơ sở dữ liệu tuần tự thu gọn thu được từ *Bước 4*, các liên kết kế tiếp được dự đoán theo mô hình CPT+ [46].

3.5.4. Độ đo đánh giá

Độ chính xác của dự đoán được xác định bằng công thức *(1.1)*:

Accuracy = |*successes*| / |*sequences*|

Trong đó

Accuracy: Độ chính xác của dự đoán.

|*successes*|: Số lượng chuỗi dự đoán thành công.

|*sequences*|: Số lượng chuỗi dự đoán.

Luận án đã sử dụng thư viện SPMF [35] để kiểm chứng độ chính xác của cơ sở dữ liệu tuần tự thu gọn bằng giải thuật PageRank với cơ sở dữ liệu tuần tự gốc.

**** Độ phức tạp của thuật toán:***

- ✓ Đối với phương pháp dự đoán truy cập Web dùng CPT+:

  Trong trường hợp xấu nhất độ phức tạp là *O(N1*Avg_n)*, với N1 là số lượng các chuỗi tuần tự, Avg_n là chiều dài trung bình các chuỗi tuần tự dự đoán truy cập Web trong cơ sở dữ liệu tuần tự

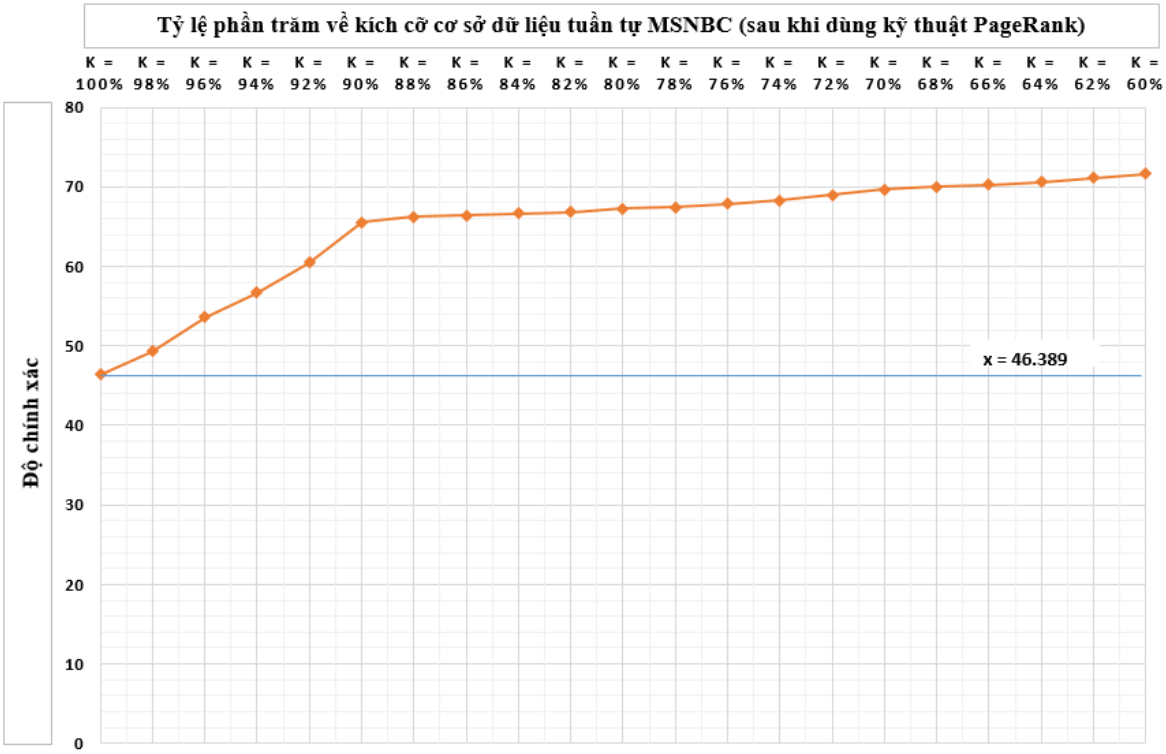
- ✓ Đối với giải pháp dự đoán truy cập Web dùng CPT+ có tích hợp Pagerank:

  Trong trường hợp xấu nhất độ phức tạp là *O(N2 *Avg_m)* , với N2 là số lượng các chuỗi tuần tự, với Avg_m là chiều dài trung bình các chuỗi tuần tự dự đoán truy cập Web trong cơ sở dữ liệu tuần tự được thu gọn bằng thuật toán PageRank.

- ✓ Do *O(N2 *Avg_m)* rất nhỏ so với *O(N1 *Avg_n)* nên độ phức tạp về không gian của giải pháp dự đoán truy cập Web dùng CPT+ có tích hợp PageRank sẽ nhỏ hơn nhiều so với phương pháp dự đoán truy cập Web dùng CPT+.

3.5.5. Các kết quả thử nghiệm

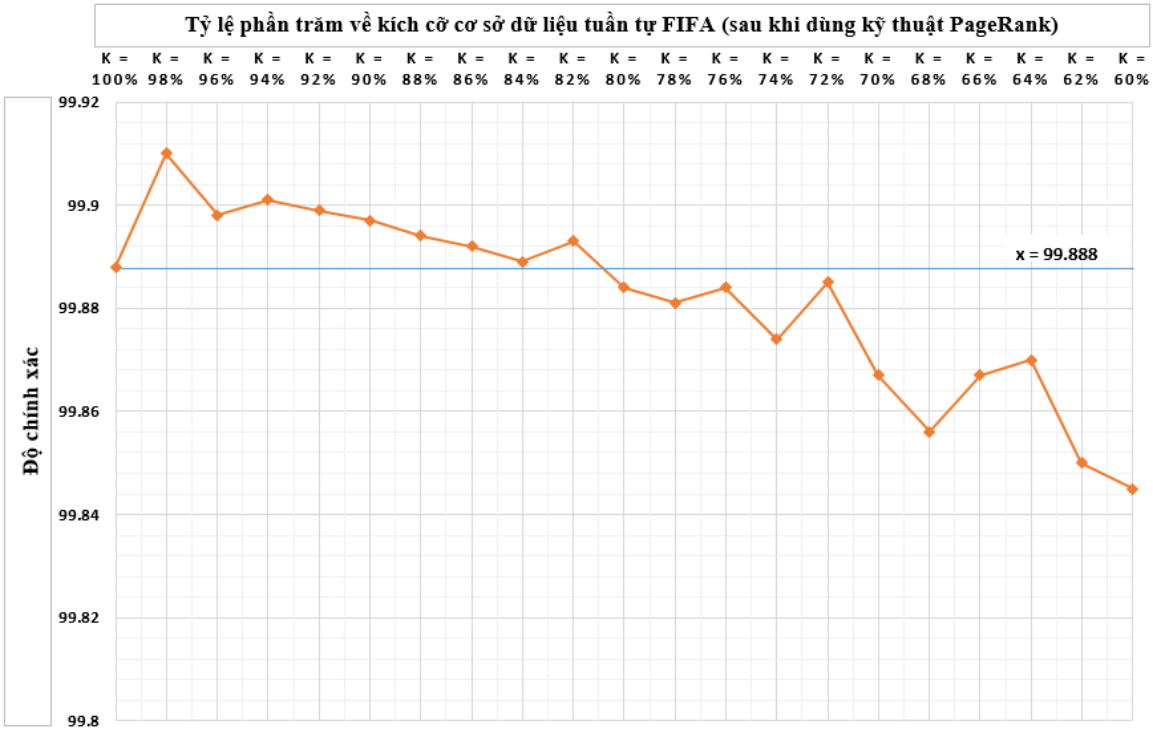
Trong thử nghiệm này, luận án đã dùng giải thuật PageRank để làm giảm kích cỡ của 3 tập dữ liệu là MSNBC, FIFA và KOSARAK. Sau đó, luận án đã kiểm tra độ chính xác trên các tập dữ liệu mới đã được thu hẹp không gian dự đoán (sau khi giảm kích cỡ của các tập dữ liệu gốc) bằng cách sử dụng thư viện SPMF [34]. Cụ thể là, nghiên cứu sinh đã chọn 21 giá trị k khác nhau, khoảng cách giữa các giá trị k là 2% (từ k = 100% giảm xuống k = 60%, với k = 100% tương ứng với kích cỡ gốc của các tập dữ liệu và k = 60% ứng với kích cỡ đã được thu hẹp còn lại 60% so với kích cỡ gốc của các tập dữ liệu). Tất cả các tập dữ liệu được thu gọn về kích cỡ có thể được tải về tại liên kết http://bit.ly/2AniqEm. Các kết quả thực nghiệm được mô tả trình tự tại ***Hình 3.4, 3.5 và 3.6***.



Hình 3.4 *So sánh độ chính xác dự đoán truy cập Web (dùng giải thuật PageRank và CPT+) trên tập dữ liệu MSNBC*

Theo mô tả của biểu đồ ở ***Hình 3.4*** , cơ sở dữ liệu tuần tự gốc MSNBC có độ chính xác dự đoán truy cập Web là 46.389 % (ứng với trường hợp k = 100%, tương ứng với đường thẳng x = 46.389). Đây là độ chính xác khá thấp đối với tập dữ liệu này. Độ chính xác đã giảm đáng để trong khoảng k ∈ [90%, 100%).

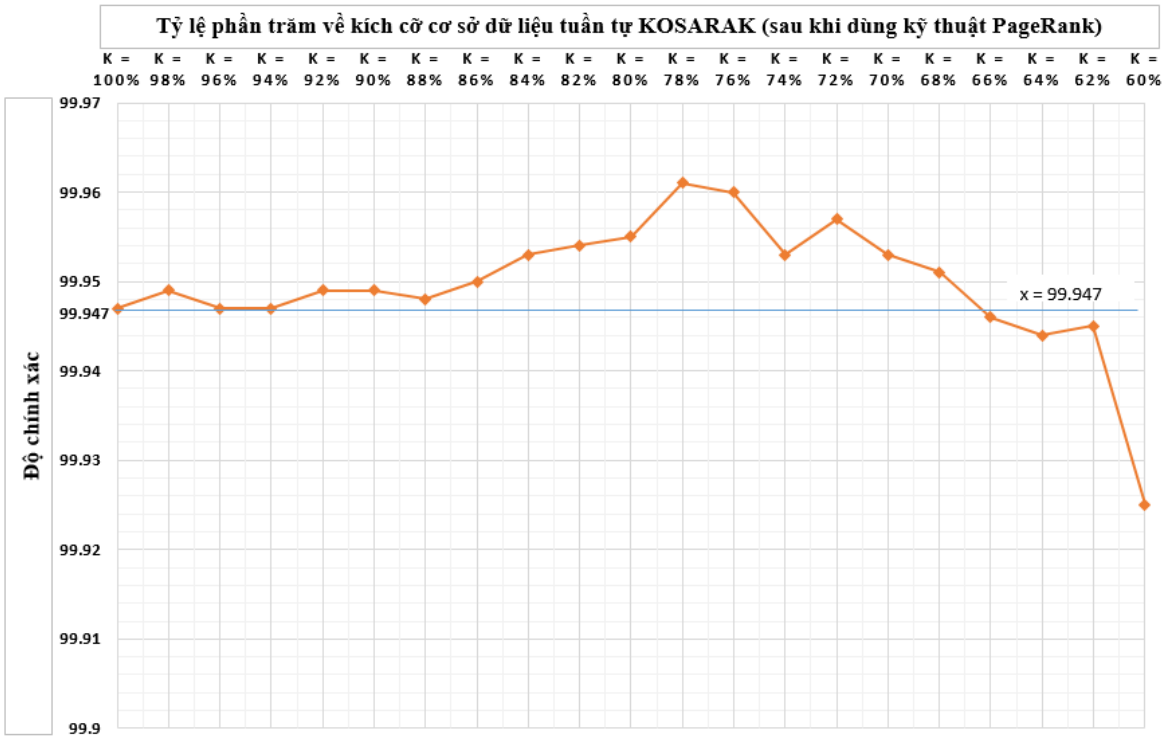
Sau đó, độ chính xác đã giảm nhẹ trong khoảng k ∈ (60%, 90%). Điều này cho thấy rằng tập dữ liệu gốc MSNBC vẫn còn chứa nhiều dữ liệu dư thừa và không có ý nghĩa. Như vậy, khi tích hợp giải thuật PageRank với CPT+ cho việc dự đoán truy cập Web trên tập dữ liệu MSNBC đã mang lại hiệu quả.



Hình 3.5 *So sánh độ chính xác dự đoán truy cập Web (dùng giải thuật PageRank và CPT+) trên tập dữ liệu FIFA*

Biểu đồ được mô tả trong ***Hình 3.5*** cho thấy tập dữ liệu gốc FIFA có độ chính xác dự đoán truy cập Web khá cao là 99.888% (với k = 100%, tương ứng với đường thẳng x = 99.888). Như được minh họa ở ***Hình 3.5***, ở khoảng k ∈ [82%, 100%), độ chính xác của các tập dữ liệu được thu gọn đều cao hơn độ chính xác của tập dữ liệu gốc FIFA. Như vậy việc tích hợp giải thuật PageRank với CPT+ để hỗ trợ dự đoán truy cập Web cho tập dữ liệu FIFA đã mang lại hiệu quả.

Tương tự, biểu đồ ở ***Hình 3.6*** cho thấy rằng tập dữ liệu gốc KOSARAK có độ chính xác dự đoán truy cập Web cũng khá cao, lên đến 99.947% (với k = 100%, tương ứng với đường thẳng x = 99.947). Mặc dù có biến độ về độ chính xác, độ chính xác của các tập dữ liệu được thu gọn luôn nhỉnh hơn độ chính xác của tập dữ liệu gốc trong khoảng k ∈ [68%, 100%). Đặc biệt khi tập dữ liệu KORASAK giảm kích cỡ đến 22% thì độ chính xác đạt đến hơn 99.96%. Như vậy việc tích hợp giải thuật PageRank với CPT+ để hỗ trợ dự đoán truy cập Web cho tập dữ liệu KORASAK là rất có ý nghĩa.



Hình 3.6 *So sánh độ chính xác dự đoán truy cập Web (dùng giải thuật PageRank và CPT+) trên tập dữ liệu KOSARAK*

Để đánh giá giải pháp rõ ràng hơn, nghiên cứu sinh đã thực hiện dự đoán truy cập Web trên 2 tập dữ liệu FIFA và KORASAK (2 tập dữ liệu này có độ chính xác về dự đoán khá cao và có mật độ dữ liệu khá dày, cụ thể là FIFA có 20,450 chuỗi dữ liệu tuần tự với 2991 trang Web khác nhau và KOSARAK có 69999 chuỗi dữ liệu tuần tự với 19986 trang Web khác nhau). Với tập dữ liệu FIFA, nghiên cứu sinh chọn k = 100% (ứng với tập dữ liệu gốc FIFA) và k = 85% (Ứng với tập dữ liệu FIFA đã giảm 15% về kích cỡ). Sau đó nghiên cứu sinh sử dụng giải thuật CPT+ để dự đoán truy cập Web và đưa ra các kết quả so sánh. Nghiên cứu sinh cũng thực hiện tương tự trên tập dữ liệu KOSARAK với k =100% và k = 68%.

Với các tập dữ liệu MSNBC, FIFA, KOSARAK, nghiên cứu sinh đã thực hiện giảm đến 50%, 15%, 30% (theo trình tự các tập dữ liệu) kích cỡ không gian dự đoán (kích cỡ của cơ sở dữ liệu tuần tự) nhưng độ chính xác của giải pháp tích hợp giải thuật PageRank với CPT+ vẫn luôn cao hơn độ chính xác của tiếp cận chỉ dùng CPT+ (kích cỡ cơ sở dữ liệu tuần tự chưa giảm kích cỡ), Với tập dữ liệu MSNBC, độ chính xác đã tăng xấp xỉ 25 %; với tập dữ liệu FIFA, độ chính xác đã tăng xấp xỉ 0.013% và với tập dữ liệu KOSARAK, độ chính xác đã tăng xấp xỉ 0.027%. Khi kiểm tra bằng K-Fold Check Validation với K=10 trên tập dữ liệu click-stream KOSARAK cho thấy có thể giảm kích cỡ cơ sở dữ liệu đến 40% để làm tăng độ chính xác trung bình cho dự đoán truy cập Web xấp xỉ 0.024%. Như vậy, đối tập dữ liệu lớn nhất trong được sử dụng trong quá trình nghiên cứu là KOSARAK thì độ chính xác đã tăng xấp xỉ 0.025% khi áp dụng tích hợp tính toán PageRank với CPT+ để dự đoán truy cập Web.

Sau đây là phần lý giải tại sao khi áp dụng tính toán PageRank tích hợp với CPT+ làm tăng độ chính xác cho dự đoán truy cập Web:

Gọi SD_{full} là cơ sở dữ liệu tuần tự gốc ban đầu.

Sau khi xử lý tính toán PageRank, cơ sở dữ liệu tuần tự được sắp xếp thành 2 phần. Phần thứ nhất: SD_{high} là tập hợp các chuỗi dữ liệu tuần tự có trung bình chỉ số PageRank cao và phần thứ hai: SD_{low} là tập hợp các chuỗi dữ liệu tuần tự có trung bình chỉ số PageRank thấp. Mối quan hệ của các tập dữ liệu này được xác định bằng công thức sau:

$$SD_{full} = SD_{high} \cup SD_{low}$$

Xét SD_{high} là tập dữ liệu có chứa những chuỗi có dạng * P_{PR}, với * là chuỗi dữ liệu tuần tự bất kỳ và dự đoán được trang P_{PR} (P_{PR} luôn theo sau các chuỗi *).

Xét những trang Web truy cập đến trang P_{PR}, khi chỉ số PageRank của trang P_{PR} cao, các trang sẽ truy cập trực tiếp đến trang P_{PR} càng nhiều (theo tính chất của PageRank). Điều đó cũng có nghĩa là sẽ xuất hiện rất nhiều những chuỗi có dạng * P_{PR}. Điều này có nghĩa là số chuỗi dự đoán thành công trang P_{PR} sẽ tăng lên. Ngược lại, khi chỉ số PageRank của trang P_{PR} thấp, các trang sẽ truy cập trực tiếp đến trang P_{PR} càng ít và là số chuỗi dự đoán thành công trang P_{PR} sẽ giảm xuống.

Khi tính toán trung bình các chỉ số PageRank trên các chuỗi dữ liệu tuần tự, chỉ số trung bình của các chuỗi tuần tự nào càng cao thì sẽ xuất hiện rất nhiều những chuỗi có dạng * P_{PR}. Điều này cũng có nghĩa là số chuỗi dự đoán thành công sẽ càng nhiều và sẽ được cộng thêm vào số chuỗi dự đoán thành công theo giải thuật CPT+. Do đó, việc tích hợp tính toán PageRank vào CPT+ là rất có ý nghĩa cho dự đoán truy cập Web.

3.6. Kết luận chương 3

Các kết quả thử nghiệm cho thấy rằng việc làm giảm kích cỡ của các tập dữ liệu (thu hẹp không gian dự đoán) mà vẫn bảo toàn độ chinh xác là rất có ý nghĩa khi thực hiện dự đoán truy cập Web trên các tập dữ liệu được thu gọn đến mức có thể (nhưng vẫn đảm bảo độ chính xác ở mức tối đa). Như được trình bày ở trên, vẫn tồn tại một số tập dữ liệu có độ chính xác dự đoán truy cập Web khá thấp (chưa

đến 50%) như trường hợp của tập dữ liệu MSNBC. Điều này không thực sự tốt cho dự đoán vì có quá nhiều dữ liệu dư thừa và không có ý nghĩa cho dự đoán. Theo kết quả thu được ở trên, độ chính xác của MSNBC đã tăng lên đến hơn 20% bằng cách áp dụng giải thuật PageRank để loại bỏ những dữ liệu xấu và không hữu ích.

Bên cạnh đó, mặc dù độ chính xác của các tập dữ liệu gốc FIFA và KOSARAK là khá cao, chúng ta vẫn có thể cải thiện thêm bằng cách dùng giải thuật PageRank để lại bỏ dữ liệu không cần thiết cho dự đoán truy cập Web. Các công trình liên quan đến nội dung này là các công trình CT7.

Như vậy, trong nghiên cứu này nghiên cứu sinh đã đề xuất một giải pháp tích hợp giải thuật PageRank với CPT+. Nghiên cứu đã được thực hiện trên 3 tập dữ liệu về truy cập Web khác nhau nhằm loại bỏ dữ liệu không cần thiết cho dự đoán để nâng cao độ chính xác cho dự đoán hành vi truy cập Web.

Trong tương lai, nghiên cứu sinh sẽ thực hiện các nghiên cứu để cải thiện hơn nữa về độ chính xác và thời gian cho dự đoán truy cập Web bằng những giải thuật mới và tốt hơn.

CHƯƠNG 4. NÂNG CAO HIỆU QUẢ VỀ THỜI GIAN KHAI PHÁ DỮ LIỆU TUẦN TỰ CHO DỰ ĐOÁN TRUY CẬP WEB

4.1. Giới thiệu

Chương 4 trình bày một giải pháp tích hợp kỹ thuật phân tích chuỗi với CPT+ để nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Dữ liệu đầu vào cho nghiên cứu là các cơ sở dữ liệu tuần tự được thu thập từ các tập dữ liệu thu thập từ người dung truy cập Web, cụ thể là các tập dữ liệu click-stream (FIFA, KOSARAK, BMS [1]) hoặc các tập dữ liệu được xây dựng từ nhật ký máy chủ của các Website (palmviewnasibel [2] , inees [3]). Tuy nhiên, để thời gian dự đoán truy cập Web được hiệu quả hơn, các cơ sở dữ liệu tuần tự này cần phải loại bỏ những chuỗi dữ liệu tuần tự mà ẩn chứa nhiều dữ liệu dư thừa và không có ý nghĩa cho dự đoán truy cập Web. Bằng giải pháp áp dụng kỹ thuật phân tích các chuỗi tuần tự trong các cơ sở dữ liệu tuần tự và kết hợp với CPT+, các cơ sở dữ liệu tuần tự thu được có kích cỡ nhỏ hơn rất nhiều để hỗ trợ cho dự đoán truy cập Web hiệu quả hơn về mặt thời gian.

4.2. Cơ sở lý luận của giải pháp

Việc xử lý dự đoán truy cập Web bằng phương pháp CPT+ tốn rất nhiều chi phí về thời gian vì thế nghiên cứu và đề xuất giải pháp tối ưu hiệu quả của giải pháp này cho dự đoán hành vi truy cập Web là rất cần thiết và hữu ích. Mặt khác, việc cải thiện độ chính xác cho dự đoán cũng cần được xem xét. Đặc biệt là tìm ra giải pháp để giảm tối đa các chuỗi dữ liệu tuần tự thừa và không có ý nghĩa cho dự đoán truy cập Web là một công việc rất quan trọng.

---

[1] https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

[2] Truy cập www.palviewnasibel.com ngày 29/9/2019

[3] Truy cập www.inees.org ngày 25/8/2017

Luận án đã đề xuất một giải pháp để giải quyết các vấn đề này. Ý tưởng của giải pháp này như sau:

Cho chuỗi tuần tự S chứa các liên kết truy cập tuần tự cần dự đoán các đối tượng kế tiếp (liên kết truy cập kế tiếp) và cơ sở dữ liệu tuần tự SDB chứa tập hợp các chuỗi tuần tự (mỗi chuỗi tuần tự trong SDB chứa các liên kết truy cập tuần tự theo thời gian), mục đích của giải pháp thu gọn cơ sở dữ liệu tuần tự là làm giảm kích cỡ của cơ sở dữ liệu tuần tự SDB ban đầu và các kết quả dự đoán các đối tượng kế tiếp không quá sai lệch so với kết quả dự đoán khi sử dụng giải pháp CPT+ [46] trên cơ sở dữ liệu tuần tự gốc ban đầu.

Thay vì sử dụng giải pháp dự đoán chuỗi tuần tự CPT+, không gian dự đoán được làm giảm kích thước bằng cách loại bỏ các chuỗi trình tự dư thừa mà không làm mất đi độ chính xác trong dự đoán.

4.3. So sánh thời gian thực thi của các tiếp cận dự đoán dữ liệu tuần tự

Nghiên cứu [47] đã chỉ ra tiếp cận CPT cho dự đoán chuỗi dữ liệu tuần tự hiệu quả hơn những phương pháp khác, cụ thể như trình bày dưới đây.

4.3.1. Các bộ dữ liệu dùng để so sánh thời gian thực thi dự đoán

Nghiên cứu [47] đã sử dụng 5 bộ dữ liệu được thu thập theo thời gian thực để nghiên cứu như sau:

BMS là bộ dữ liệu phổ biến trong khai phá dữ liệu, đặc biệt là khai phá luật kết hợp [118]. Bộ dữ liệu này đã được thu thập từ các truy cập của người dùng trên một trang Web thương mại điện tử, được mã số thành các chuỗi dữ liệu có dạng số nguyên mà biểu diễn cho các trang Web.

FIFA được thu thập từ các truy cập Web được ghi nhận trên trang Web FIFA World Cup 1998. Dữ liệu đã được mã hóa cho phù hợp với yêu cầu nghiên cứu dự đoán chuỗi dữ liệu tuần tự.

SIGN là một bộ dữ liệu dày đặc với các chuỗi dài, chứa 730 chuỗi những lời nói ngôn ngữ ký hiệu được thu thập từ video [87].

KOSARAK là bộ dữ liệu chứa các truy cập Web của người dùng trên một cổng thông tin tin tức Hungary (fimi.ua.ac.be/data).

BIBLE là bộ dữ liệu được thu thập và mã hóa từ tập sách Kitô giáo. Nhiệm vụ dự đoán dữ liệu bao gồm việc dự đoán ký tự tiếp theo trong chuỗi ký tự được đưa ra. Cuốn sách được chia thành các câu trong đó mỗi câu là một chuỗi. Tập dữ liệu này rất thú vị vì nó có một bảng chữ cái nhỏ chỉ với 75 ký tự riêng biệt và dựa trên ngôn ngữ tự nhiên [47].

Trong các tập dữ liệu được nêu ở trên có 3 tập dữ liệu được thu thập từ các truy cập Web là BMS, FIFA và KOSARAK.

4.3.2. So sánh thời gian của các tiếp cận dự đoán dữ liệu tuần tự

Trong phần nghiên cứu này, CPT được so sánh với những tiếp cận dự đoán dữ liệu tuần tự phổ biến khác như DG [84], PPM [23], AKOM [93] theo ***Bảng 4.1***.

Bảng 4.1 *So sánh thời gian thực thi của CPT so với các tiếp cận khác [47]*

Tập dữ liệu	**Thời gian huấn luyện (Training)** *(seconds)*				**Thời gian kiểm thử (Testing)** *(seconds)*			
	DG	CPT	PPM	AKOM	DG	CPT	PPM	AKOM
BMS	0.076	0.018	0.01	0.356	0.004	0.352	**0.001**	0.004
FIFA	3.032	0.153	**0.095**	12.347	0.301	0.146	**0.006**	0.085
KOSARAK	9.697	0.741	**0.173**	6.051	0.042	1.533	0.018	**0.011**

Về thời gian huấn luyện (training time), kết quả nghiên cứu của [47] theo ***Bảng 4.1*** cho thấy với các bộ dữ liệu truy cập Web như BMS, FIFA, thời gian thực thi của CPT chỉ chậm hơn PPM.

Về thời gian dự đoán, trên bộ dữ liệu BMS và KOSARAK, CPT thực thi chậm nhất; trên bộ dữ liệu FIFA, CPT nhanh hơn gần gấp 3 lần so với DG nhưng chậm hơn so với PPM và AKOM; trên bộ dữ liệu KOSARAK.

Như vậy, CPT đã thực thi chậm hơn so với các tiếp cận dự đoán dữ liệu tuần tự khác. Tuy nhiên, theo bài báo [46], độ chính xác của CPT vượt trội hơn so với các tiếp cận này như minh họa ở ***Bảng 1.3***.

Hơn nữa, một tiếp cận cải tiến của CPT là CPT+ [46] đã cho thấy thế mạnh vượt trội của mình về thời gian thực thi (nhanh gần 5 lần) và độ chính xác (5%) so với tiếp cận CPT, xem chi tiết trong ***Bảng 3.3***. Bên cạnh đó, so với các tiếp cận phổ biến về dự đoán chuỗi dữ liệu tuần tự khác, CPT+ đã cho thấy độ chính xác khá cao so với các tiếp cận như CPT [47], All-K-Order-Markov(AKOM) [93], Dependancy Graph (DG) [84] , LZ78 [121], PPM [23], Transition Directed Acyclic Graph(TDAG) [72]. Đặc biệt, các kết quả thực nghiệm của [46] trên các bộ dữ liệu truy cập Web như BMS, FIFA, KOSARAK đã cho thấy rằng CPT+ là giải pháp tốt nhất. Mặc dù vậy, trong một trường hợp riêng lẻ, cụ thể là trên bộ dữ liệu MSNBC (một truy cập Web được thu thập từ kho khai phá dữ liệu UCI https://archive.ics.uci.edu/ml) thì độ chính xác của CPT+ hơi kém hơn so với tiếp cận CPT. Tuy nhiên, kích cỡ dữ liệu là của MSNBC chỉ xấp xỉ là 50% so với FIFA và chỉ khoảng 30% so với KOSARAK. Ngoài ra, FIFA và KOSARAK là hai bộ dữ liệu tin cậy hơn vì được sử dụng phổ biến hơn so vớ MSNBC.

Từ phân tích trên cho thấy CPT+ là một tiếp cận phù hợp nhất trong thời điểm này. Tuy nhiên việc tiếp tục nâng cao hiệu quả về thời gian của tiếp cận CPT+ là rất cần thiết vì thời gian sẽ chậm dần khi tăng dần kích cỡ của không gian dự đoán (chẳng hạn tăng về kích cơ của cơ sở dữ liệu tuần tự).

Phần tiếp theo trình bày sẽ đề xuất chi tiết một giải pháp để nâng cao hiệu quả của dự đoán truy cập Web.

4.4. Giải pháp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web với CPT+

4.4.1. Cơ sở lý luận của giải pháp

Chúng ta đều biết rằng dữ liệu thừa hay gây nhiễu thông tin trong truy cập Web là khá lớn. Trong trường hợp của giải pháp này dữ liệu thừa mà nghiên cứu sinh đề cập đến là ngữ cảnh khi người dùng chỉ truy cập vào một trang Pz mà không tiếp tục thực hiện bất kỳ truy cập nào khác. Như vậy trang Pz mà người dùng truy cập không thể dẫn đến một trang khác (dự đoán một trang khác). Cụ thể trong giải pháp này trang Px được nêu chính là trang tận cùng của một chuỗi dữ liệu tuần tự trong cơ sở liệu tuần tự cần khai phá.

4.4.2. Giải thuật nâng cao hiệu quả về thời gian dự đoán truy cập Web

Mô tả giải thuật nâng cao hiệu quả về thời gian truy cập Web:

Dữ liệu nhập vào:

+ arr_sequence: Mảng chứa các chuỗi tuần tự trong cơ sở dữ liệu tuần tự

+ arr_query: Mảng chứa các phần tử trong chuỗi dữ liệu cần dự đoán phần tử kế tiếp

Dữ liệu thu được: Cơ sở dữ liệu tuần tự đã được thu gọn

Chi tiết mã giả (Pseudo Code) của Bước 2 như sau:

```
1. //Tìm các chuỗi tuần tự có chứa chuỗi cần dự đoán phần tử kế tiếp
2. Cấp phát mảng chuỗi seq có n  phần tử
3. k := 0; //k: số lượng các các phần tử trong chuỗi dữ liệu cần dự đoán
4. str_contain_query = " "; SD_OK = null;
5. // str_contain_query  là chuỗi chứa chuỗi tuần tự cần dự đoán
6. For i = 0 to (k-1) do
7. If (arr_sequence[i] có chứa ít nhất một phần tử thuộc query) Then
```

```
8. Begin
9. If (query ⊆ arr_contain_query[i] and it is not at  the last position of ⊈
10. arr_contain_query[i] Or (query ⊆ arr_contain_query[i] and it is at the
11. last position of arr_contain_query[i]  And Card{query ⊆
12. arr_contain_query[i]} > 1)) Then
13.     Begin
14.         SD_OK + = arr_contain_query[i] // Chuỗi tuần tự hợp lệ được chọn
15.     End
16. End
```

*** Đánh giá độ phức tạp của giải thuật**

Độ phức tạp của giải thuật được tính toán dựa theo các chi phí thời gian thực hiện các mã nguồn *[Source 6]* ở phần PHỤ LỤC 2.

Xét *arr_sequence* là mảng răng cưa (jagged array) với các phần tử là các chuỗi tuần tự, N = *arr_sequence.length* là số lượng các chuỗi tuần tự trong cơ sở dữ liệu tuần tự và *k* là số lượng các chuỗi tuần tự mà có chứa chuỗi tuần tự cần dự đoán truy cập kế tiếp.

- Xét [Source 6] (PHỤ LỤC 2):
  - ✓ Dòng lệnh 5 và dòng lệnh 6 tốn chi phí thời gian là O(1) do đó dòng lệnh điều kiện 4 có chi phí thời gian là O(1) và dòng lệnh điều kiện 2 có chi phí thời gian là O(1)
  - ✓ Vòng lặp 1 thực hiện N lần, mỗi lần tốn chi phí thời gian là O(1) do đó độ phức tạp để chương trình thực hiện vòng lặp 1 là O(N)
  - ✓ Các dòng lệnh 9, 10, 11 tốn chi phí thời gian là O(1)
  - ✓ Các dòng lệnh 17, 18 tốn chi phí thời gian là O(1) do đó dòng lệnh điều kiện 13 tốn chi phí thời gian là O(1).
  - ✓ Vòng lặp 12 thực hiện k lần, mỗi lần tốn chi phí O(1) do đó độ phức tạp để chương trình thực hiện vòng lặp 12 là O(k)

Như vậy *[Source 6]* có độ phức tạp là O(Max(O(N), O(k)).

Mà $k \ll N$, do đó *[Source 6]* có độ phức tạp là O(N).

4.5. Các kết quả thử nghiệm nâng cao hiệu năng thời gian thực thi dự đoán truy cập Web

4.5.1 Mục tiêu

Phần này trình bày các kết quả thử nghiệm nâng cao hiệu năng thời gian dự đoán truy cập Web trên 3 tập dữ liệu Click-stream và 2 tập dữ liệu Weblog bằng phương pháp phân tích chuỗi dự đoán đã trình bày ở phần 4.4.

4.5.2. Dữ liệu

Đối các tập dữ liệu click-stream, các cơ sở dữ liệu tuần tự được sử dụng trong thử nghiệm: *FIFA* [1], *KOSARAK* [10], *BMS* [10].

Bảng 4.2 *Các tập dữ liệu click-stream được thử nghiệm*

Tập dữ liệu	**Số lượng chuỗi tuần tự**
FIFA	20540
KORARAK	69999
BMS	77512

Đối với các tập dữ liệu thu thập từ Weblog, các cơ sở dữ liệu tuần tự được sử dụng trong thử nghiệm: *palmviewsanibel* [2] , *inees* [3].

Bảng 4.3 *Các tập dữ liệu Weblog được thử nghiệm*

Tập dữ liệu	**Số lượng chuỗi tuần tự**
palmviewsanibel	4967 (được chuẩn hóa từ 5282543 mẫu tin Weblog)
Inees	995 (được chuẩn hóa từ 1522983 mẫu tin Weblog)

4.5.3. Phương pháp

Nghiên cứu sinh đã phát triển [CT2] để làm giảm kích cỡ của cơ sở dữ liệu tuần tự ban đầu nhằm làm tăng hiệu quả về thời gian xử lý cho dự đoán truy cập Web. Chi tiết giải pháp đề xuất được thực hiện như sau:

Dữ liệu nhập:

- ✓ Chuỗi tuần tự cần dự đoán *S_query*
- ✓ Cơ sở dữ liệu tuần tự *SDB*

Xử lý:

[1] https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php , truy cập ngày 12/12/2018

[2] Truy cập www.palviewnasibel.com ngày 29/9/2019

[3] Truy cập www.inees.org ngày 25/8/2017

Khởi tạo thời gian thực hiện việc xử lý. Gọi thời gian khởi tạo này là *T1*

- Bước 1:

Xét tất cả các chuỗi tuần tự *S* thuộc *SDB*, tiến hành loại bỏ các chuỗi tuần tự *S* nào mà không chứa ít nhất một phần tử thuộc *S_query*. Gọi cơ sở dữ liệu mới thu được là *SDB1* và kích cỡ tương ứng là *SDB1_size.*

- Bước 2:

Tiếp tục thực hiện trên *SDB1*: Loại bỏ các chuỗi tuần tự có chứa duy nhất chuỗi tuần tự *S_query* nằm ở vị trí tận cùng của các chuỗi tuần tự trong *SDB1* vì những chuỗi tuần tự này không có ý nghĩa để dự đoán phần tử kế tiếp. Gọi cơ sở dữ liệu mới thu được sau khi thực hiện bước này là *SDB2* và kích cỡ tương ứng là *SDB2_size.*

- Bước 3:

Áp dụng giải thuật CPT+ để dự đoán truy cập Web trên cơ sở dữ liệu *SD2.* Ghi nhận thời gian thực hiện hai bước trên (T_1)

Tính độ đo Acc_1 [48].

Kết quả thu được:

- ✓ Kích cỡ cơ sở dữ liệu tuần tự *SD2_size*.
- ✓ Độ đo Accuracy: Acc_1.
- ✓ Thời gian thực thi: T_1.

Với tiếp cận truyền thống, chỉ sử dụng CPT+ cho dự đoán truy cập Web, Bước 2 sẽ không được thực hiện. Kết quả thu được như sau:

- ✓ Kích cỡ cơ sở dữ liệu tuần tự *SD_size*.
- ✓ Độ đo Accuracy: *Acc*.
- ✓ Thời gian thực thi: *T*.

Vấn đề được đặt ra :

+ Thời gian thực thi T_1 có nhanh hơn Thời gian thực thi *T* đáng kể hay không?

+ Độ chính xác Acc_1 có tương đương hay cao hơn độ chính xác *Acc*?

4.5.4. Các độ đo đánh giá

Nghiên cứu sinh đã sử dụng thư viện SPMF [35] để kiểm chứng độ chính xác của cơ sở dữ liệu tuần tự thu gọn (có tích hợp giải pháp phân tích chuỗi) so với cơ sở dữ liệu tuần tự gốc.

❖ Độ đo đánh giá về độ chính xác:

Áp dụng công thức *(1.1)* cho Acc và Acc_1

Nếu $Acc_1 \geq Acc$: Dự đoán hiệu quả, ngược lại thì dự đoán không hiệu quả

Độ đo đánh giá về thời gian:

❖ Độ đo đánh giá về thời gian:

Nếu T_1 nhỏ hơn rất nhiều so với T : Dự đoán hiệu quả về thời gian, ngược lại thì dự đoán không hiệu quả về thời gian.

4.5.5. Kết quả thử nghiệm và phân tích

4.5.5.1. Kết quả thử nghiệm trên tập dữ liệu FIFA

Kiểm định thời gian thực thi dự đoán, độ đo Accuracy sử dụng phương pháp kiểm định Paired T-Test với 100 chuỗi dự đoán khác nhau với độ tin cậy 99% trên tập dữ liệu FIFA.

Bảng 4.4 *Kiểm định Paired T-Test cho thời gian thực thi dự đoán và độ chính xác trên tập dữ liệu FIFA*

Phương pháp	**Thời gian Thực thi trung bình (Mean)** *(milliseconds)*	**Giá trị thống kê (theo thời gian) (pValue)**	**Độ chính xác trung bình (Mean)** *(%)*	**Giá trị thống kê (theo độ chính xác) (pValue)**
CPT+ truyền thống	18763.55	0.000 (< 0.01)	99.07	0.271 (> 0.01)

CPT + cải tiến (tích hợp phân tích chuỗi)	6092.15		98.91	

Bảng 4.4 trình bày kết quả kiểm định Paired T-Test khi so sánh thời gian thực thi dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu FIFA. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 30 lần. Hơn nữa giá trị thống kê pValue = 0.000 < 0.01 (độ tin cậy 99%) cho thấy rằng có sự khác biệt rất rõ ràng về thời gian thực thi giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với thời gian thực thi theo phương pháp truyền thống (chỉ áp dụng CPT+). ***Bảng 4.4*** cũng thể hiện kết quả kiểm định Paired T-Test khi so sánh độ chính xác dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu FIFA. Kết quả cho thấy độ chính xác của hai phương pháp là tương đương nhau. Hơn nữa giá trị thống kê pValue = 0.271 > 0.01 (độ tin cậy 99%) cho thấy rằng không có sự khác biệt về độ chính xác dự đoán giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với độ chính xác dự đoán theo phương pháp truyền thống (chỉ áp dụng CPT+).

4.5.5.2. Kết quả thử nghiệm trên tập dữ liệu KOSARAK

Kiểm định thời gian thực thi dự đoán, độ đo Accuracy sử dụng phương pháp kiểm định Paired T-Test với 100 chuỗi dự đoán khác nhau với độ tin cậy 99% trên tập dữ liệu KOSARAK.

Bảng 4.5 *Kiểm định Paired T-Test thời gian dự đoán và độ chính xác trên tập dữ liệu KOSARAK*

Phương pháp	**Thời gian Thực thi**	**Giá trị thống kê**	**Độ chính xác trung bình**	**Giá trị thống kê**

	trung bình (Mean) *(milliseconds)*	**(theo thời gian) (pValue)**	**(Mean)** *(%)*	**(theo độ chính xác) (pValue)**
CPT+ truyền thống	28166.08	0.000 (< 0.01)	99.59	0.000 (< 0.01)
CPT + cải tiến (tích hợp phân tích chuỗi)	912.17		99.84	

Bảng 4.5 trình bày kết quả kiểm định Paired T-Test khi so sánh thời gian thực thi dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu KOSARAK. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 30 lần. Hơn nữa giá trị thống kê pValue = 0.000 < 0.01 (độ tin cậy 99%) cho thấy rằng có sự khác biệt rất rõ ràng về thời gian thực thi giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với thời gian thực thi theo phương pháp truyền thống (chỉ áp dụng CPT+). ***Bảng 4.5*** cũng thể hiện trình bày kết quả kiểm định Paired T-Test khi so sánh độ chính xác dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu KOSARAK. Kết quả cho thấy độ chính xác của hai phương pháp là tương đương nhau. Hơn nữa giá trị thống kê *pValue = 0.000 < 0.01* (độ tin cậy 99%) cho thấy rằng có sự khác biệt về độ chính xác dự đoán giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với độ chính xác dự đoán theo phương pháp truyền thống (chỉ áp dụng CPT+). Cụ thể là độ chính xác của phương pháp cải tiến tốt hơn so với độ chính xác phương pháp truyền thống.

4.5.5.3. Kết quả thử nghiệm trên tập dữ liệu BMS

Kiểm định thời gian thực thi dự đoán, độ đo Accuracy sử dụng phương pháp kiểm định Paired T-Test với 100 chuỗi dự đoán khác nhau với độ tin cậy 99% trên tập dữ liệu BMS.

Bảng 4.6 *Kiểm định Paired T-Test thời gian dự đoán và độ chính xác trên tập dữ liệu BMS*

Phương pháp	**Thời gian Thực thi trung bình (Mean)** *(milliseconds)*	**Giá trị thống kê (theo thời gian) (pValue)**	**Độ chính xác trung bình (Mean)** *(%)*	**Giá trị thống kê (theo độ chính xác) (pValue)**
CPT+ truyền thống	36361.56	0.000 (< 0.01)	100	0.320 (> 0.01)
CPT + cải tiến (tích hợp phân tích chuỗi)	351.39		99.99	

Bảng 4.6 trình bày kết quả kiểm định Paired T-Test khi so sánh thời gian thực thi dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu BMS. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 100 lần. Hơn nữa giá trị thống kê *pValue = 0.000 < 0.01* (độ tin cậy 99%) cho thấy rằng có sự khác biệt rất rõ ràng về thời gian thực thi giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và

CPT+) so với thời gian thực thi theo phương pháp truyền thống (chỉ áp dụng CPT+). ***Bảng 4.6*** cũng trình bày kết quả kiểm định Paired T-Test khi so sánh độ chính xác dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu BMS. Kết quả cho thấy độ chính xác của hai phương pháp là tương đương nhau. Hơn nữa giá trị thống kê *pValue = 0.320 > 0.01* (độ tin cậy 99%) cho thấy rằng không có sự khác biệt về độ chính xác dự đoán giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với độ chính xác dự đoán theo phương pháp truyền thống (chỉ áp dụng CPT+).

4.5.2.4. Kết quả thử nghiệm trên tập dữ liệu pamviewsanibel

Kiểm định thời gian thực thi dự đoán, độ đo Accuracy sử dụng phương pháp kiểm định Paired T-Test với từng chuỗi dự đoán trong **Phụ lục 3** với độ tin cậy 99% trên tập dữ liệu *palmviewsanible*.

Bảng 4.7 *Kiểm định Paired T-Test thời gian dự đoán và độ chính xác trên tập dữ liệu palmviewsanibel*

Phương pháp	**Thời gian Thực thi trung bình (Mean)** *(milliseconds)*	**Giá trị thống kê (theo thời gian) (pValue)**	**Độ chính xác trung bình (Mean)** *(%)*
CPT+ truyền thống	397.89	0.000 (< 0.01)	100
CPT + cải tiến	146.74		100

(tích hợp phân tích chuỗi)			

Bảng 4.7 trình bày kết quả kiểm định Paired T-Test khi so sánh thời gian thực thi dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu *palmviewsanibel* . Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn khoảng 2.7 lần. Hơn nữa giá trị thống kê *pValue = 0.000 < 0.01* (độ tin cậy 99%) cho thấy rằng có sự khác biệt rất rõ ràng về thời gian thực thi giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với thời gian thực thi theo phương pháp truyền thống (chỉ áp dụng CPT+). ***Bảng 4.7*** cũng trình bày kết quả kiểm định Paired T-Test khi so sánh độ chính xác dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu *palmviewsanibel*. Kết quả cho thấy độ chính xác dự đoán giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với độ chính xác dự đoán theo phương pháp truyền thống (chỉ áp dụng CPT+) là không thay đổi.

4.5.2.5. Kết quả thử nghiệm trên tập dữ liệu inees

Kiểm định thời gian thực thi dự đoán, độ đo Accuracy sử dụng phương pháp kiểm định Paired T-Test với 100 chuỗi dự đoán khác nhau với độ tin cậy 99% trên tập dữ liệu *inees*.

Bảng 4.8 *Kiểm định Paired T-Test thời gian dự đoán và độ chính xác trên tập dữ liệu inees*

Phương pháp	**Thời gian Thực thi trung bình (Mean)** *(milliseconds)*	**Giá trị thống kê (theo thời gian) (pValue)**	**Độ chính xác trung bình (Mean)** *(%)*

CPT+ truyền thống	901.58	0.000 (< 0.01)	100
CPT + cải tiến (tích hợp phân tích chuỗi)	452.50		100

Bảng 4.8 trình bày kết quả kiểm định Paired T-Test khi so sánh thời gian thực thi dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu *inees*. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn gần 2 lần. Hơn nữa giá trị thống kê *pValue = 0.000 < 0.01* (độ tin cậy 99%) cho thấy rằng có sự khác biệt rất rõ ràng về thời gian thực thi giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với thời gian thực thi theo phương pháp truyền thống (chỉ áp dụng CPT+). ***Bảng 4.8*** cũng trình bày kết quả kiểm định Paired T-Test khi so sánh độ chính xác dự đoán truyền thống so với thời gian dự đoán cải tiến trên cơ sở dữ liệu *inees*. Kết quả cho thấy độ chính xác dự đoán giải pháp cải tiến (áp dụng kỹ thuật xử lý chuỗi và CPT+) so với độ chính xác dự đoán theo phương pháp truyền thống (chỉ áp dụng CPT+) là không thay đổi.

Tóm lại, giải pháp tích hợp phân tích chuỗi dữ liệu dự đoán vào CPT+ đã hiệu quả hơn về thời gian thực thi so với phương pháp dự đoán chỉ dùng CPT+ (không tích hợp phân tích chuỗi dự đoán). Bên cạnh đó, độ chính xác dự đoán cũng không có sự khác biệt đáng kể.

Các thực nghiệm cho thấy dự đoán trên dữ liệu trên các tập dữ liệu Click-stream cho thấy hiệu quả về thời gian hơn so với các tập dữ liệu được thu thập từ Web log.

4.6. Kết luận chương 4

Chương này đã trình bày đề xuất một giải pháp để nâng cao hiệu quả về thời gian thực thi dự đoán. Cụ thể là dự đoán các liên kết truy cập kế tiếp của các chuỗi tuần tự các liên kết truy cập tuần tự. Nghiên cứu sinh đã thử nghiệm giải pháp dự đoán chuỗi tuần tự cải tiến bằng cách tích hợp phương pháp phân tích chuỗi dự đoán với phương pháp CPT+: Bằng cách thức này, các chuỗi tuần tự dư thừa, không có ý nghĩa cho dự đoán bị loại bỏ, điều này cũng làm giảm kích cỡ không gian dự đoán của cơ sở dữ liệu tuần để dự đoán được hiệu quả hơn. Nghiên cứu sinh đã thử nghiệm trên 3 tập dữ liệu click-stream khác nhau, 2 tập dữ liệu thu thập từ Weblog và thu được các kết quả hiệu năng thời gian của các tập dữ liệu click-stream tốt hơn so với các tập dữ liệu Weblog khi dùng phương pháp phân tích chuỗi dự đoán. Bên cạnh đó, hai công trình liên quan đến luận án cũng đã được xuất bản [CT1, CT4].

CHƯƠNG 5. TÍCH HỢP NÂNG CAO ĐỘ CHÍNH XÁC VÀ NÂNG CAO HIỆU QUẢ VỀ THỜI GIAN KHAI PHÁ DỮ LIỆU TUẦN TỰ CHO DỰ ĐOÁN TRUY CẬP WEB

5.1. Giới thiệu

Các Chương 3 và Chương 4 trong luận án đã trình bày từng giải pháp riêng lẻ để khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Nếu Chương 3 của luận án trình bày hợp giải pháp nâng cao về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web thì Chương 4 trình bày giải pháp nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Chương 5 trình bày giải pháp kết hợp các giải pháp này, nghĩa là giải pháp vừa nâng cao về độ chính xác vừa nâng cao hiệu năng về thời gian cho dự đoán truy cập Web tập dữ liệu click-stream lớn nhất mà luận án nghiên cứu là KOSARAK với số lượng các chuỗi dữ liệu tuần tự là 100,000.

Giải pháp đề xuất ở Chương 5 sẽ được trình bày theo các giai đoạn sau:

(1) Giai đoạn 1: Dùng phương pháp K-Fold Crosss Validation để chia tập dữ liệu quan sát thành 10 phần dữ liệu xấp xỉ bằng nhau (K = 10). Trong mỗi phần đó chia thành 2 nhóm nhỏ với dữ liệu ngẫu nhiên. Nhóm thứ nhất gồm có 90% dữ liệu để thực hiện việc huấn luyện, 10% dùng để kiểm thử dự đoán.

(2) Giai đoạn 2: Với từng phần dữ liệu, mỗi nhóm dữ liệu huấn luyện tương ứng sẽ áp dụng giải pháp nâng cao độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web: Cụ thể là giảm kích cỡ và kiểm tra độ chính xác dự đoán của các cơ sở dữ liệu tuần tự được thu gọn.

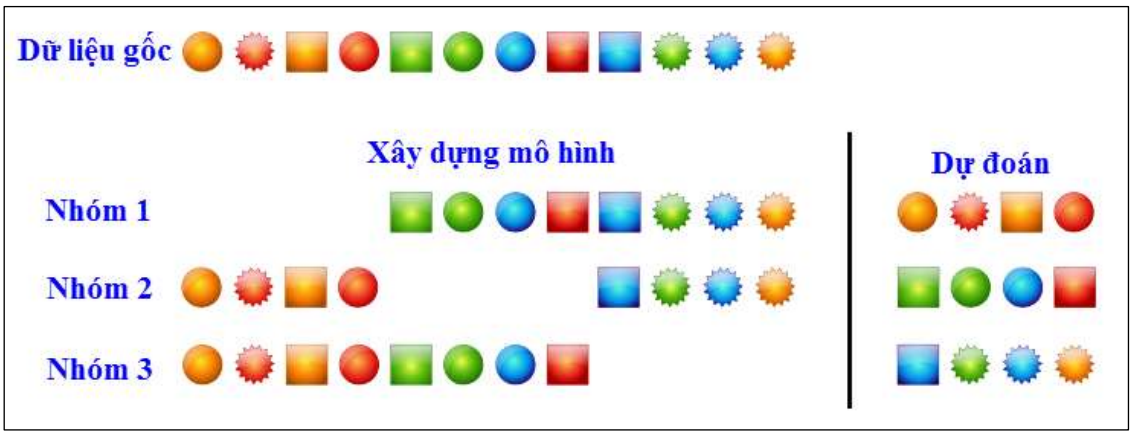
(3) Giai đoạn 3: Áp dụng giải pháp nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web cho dự đoán truy cập Web cho các cơ sở dữ liệu tuần tự đã được thu gọn ở Giai đoạn 2.

5.2. Tích hợp phương pháp K-Fold Cross Validation cho giải pháp nâng cao độ chính xác khai phá dữ liệu cho dự đoán truy cập Web

5.2.1 Phương pháp K-Fold Cross Validation

Phương pháp K-Fold Cross Validation [66] chia tập hợp các quan sát thành K nhóm, xấp xỉ với kích thước bằng nhau [58]. K thường được chọn là 5 hoặc 10 và khi K trở nên lớn hơn, sự khác biệt về kích thước giữa tập huấn luyện và các tập con lấy mẫu lại sẽ nhỏ hơn, khi sự khác biệt này càng giảm, độ lệch của kỹ thuật càng thấp [67]. Dữ liệu được huấn luyện và kiểm thử K lần, mỗi lần t {1, 2, ..., k}, được huấn luyện trên tập $D \backslash D_t$ và kiểm thử trên D_t (D là tập dữ liệu gốc và D_t là tập dữ liệu kiểm thử) [66]. Ước lượng độ chính xác của cross-validation là tổng cộng số phân loại đúng chia cho số thực thể trong tập dữ liệu gốc.

Mục đích của K-Fold Cross Validation chủ yếu được sử dụng trong Machine Learning để ước tính khả năng của mô hình học máy trên dữ liệu không nhìn thấy.



Hình 5.1 *Minh họa K-Fold Cross Validation với K = 3*

Hình 5.1 trình bày phương pháp K-Fold Cross Validation, dữ liệu gốc được chia thành 3 nhóm. Trong mỗi nhóm chia thành 2 tập con, một tập dữ liệu dùng để huấn luyện, tập còn lại dùng để dự đoán.

5.2.2. Xây dựng các tập dữ liệu huấn luyện và nâng cao độ chính xác

5.2.2.1. Mục tiêu

Việc thực hiện kiểm tra chéo bằng phương pháp K-Fold Check Validation nhằm để tạo ra 10 bộ cơ sở dữ liệu tuần tự một cách ngẫu nhiên từ cơ sở dữ liệu gốc.

Thực hiện điều này giúp chúng ta khai phá dữ liệu được khách quan và mang tính tin cậy hơn.

5.2.2.2. Dữ liệu

Bộ dữ liệu được chọn là Kosarak, đây là cơ sở dữ liệu tuần tự lớn nhất đã được giới thiệu trong các chương trước. Kích cỡ của cơ sở dữ liệu tuần tự được sử dụng trong Chương này là 100,000 chuỗi dữ liệu tuần tự [1]

5.2.2.3. Phương pháp

Việc xây dựng các tập huấn luyện và kiểm thử dự đoán được thực hiện 10 lần:

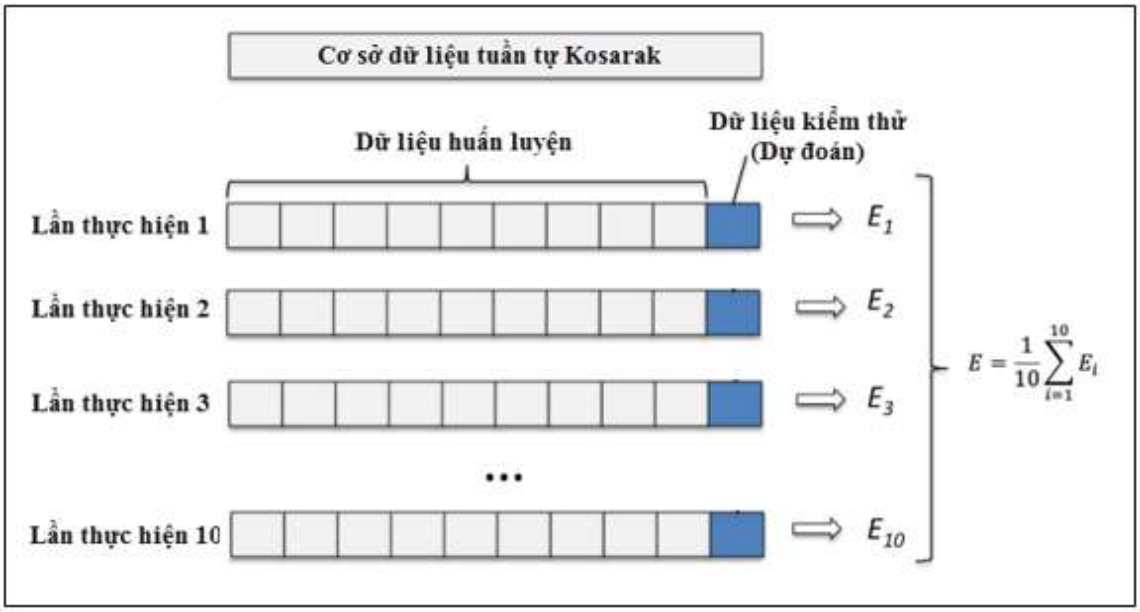
Lần thực hiện thứ nhất: Thực hiện đảo ngẫu nhiên các chuỗi tuần tự trong cơ sở dữ liệu Kosarak (100,000 dòng). Sau đó, cơ sở dữ liệu tuần tự đã tạo ra được chia thành 2 tập con:

90% về kích cỡ dữ liệu của cơ sở dữ liệu tuần tự Kosarak thu được cơ sở dữ liệu tuần tự huấn luyện *D_Training_1* (90,000 dòng), 10% dữ liệu còn lại của cơ sở dữ liệu tuần tự Kosarak là tập dữ liệu kiểm thử dự đoán, kí hiệu là *D_Testing_1* (10,000 dòng).

Lần thực hiện thứ hai: Thu được cơ sở dữ liệu tuần tự huấn luyện *D_Training_2* và *D_Testing_2*.

Sau 10 lần thực hiện, các cặp dữ liệu thu được lần lượt là (D_Training_1, D_Testing_1), (D_Training_2, D_Testing_2), …, (D_Training_10, D_Testing_10)

Hình 5.2 minh họa quá trình thực hiện để xây dựng các tập dữ liệu huấn luyện và các tập dữ liệu kiểm thử dự đoán này.



Hình 5.2 *Xây dựng các tập dữ liệu huấn luyện và kiểm thử dự đoán.*

5.2.2.4. Kết quả thực nghiệm và phân tích

Sau khi tạo ra các 10 bộ dữ liệu theo phương pháp trên, nghiên cứu sinh tiến hành lấy các 10 tập huấn luyện (có kích cỡ là 90,000 dòng) của 10 bộ dữ liệu này để thực hiện giải pháp rút gọn các chuỗi dữ liệu thừa bằng giải thuật PageRank như đã đề xuất ở Chương 3, các cơ sở dữ liệu tuần tự với độ chính xác tương ứng được tạo ra như minh họa ở ***Bảng 5.1***. Trong đó Ri là độ chính xác của các cơ sở dữ liệu tuần tự thu gọn trong lần thực hiện K-Fold Check Validation thứ i. Theo ***Bảng 5.1***, các giá trị 100, 98, 96 ...58, 56 lần lượt là kích cỡ (tính theo phần trăm) của cơ sở dữ liệu thu gọn so với cơ sở dữ liệu huấn luyện.

Kết quả thực nghiệm cho thấy rằng khi áp dụng giải pháp PageRank để giảm dần kích cỡ tập dữ liệu huấn luyện lần lượt từ 2%, 4%, 6%, ...34% (ứng với các tập dữ liệu thu gọn là 98%, 96%, 94%, ...66%), độ chính xác (được tính theo công thức

[1] Trích từ http://fimi.uantwerpen.be/data/kosarak.dat ngày 02/06/2020

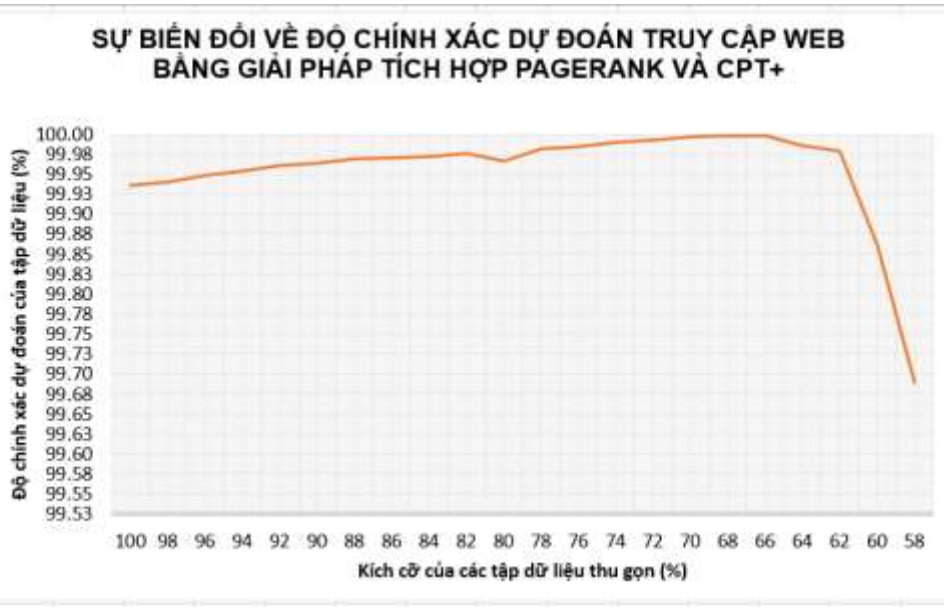
(1.1)) độ chính xác của cơ sở dữ liệu huấn luyện ban đầu. Quá trình xây dựng các cơ sở dữ liệu tuần tự huấn luyện thu gọn được thực hiện trong thời gian sắp xỉ 18 ngày (440 giờ) vì bộ dữ liệu khá lớn (100,000 dòng) và số lượng nút trong đồ thị có hướng (mô tả trong Chương 3) cũng không nhỏ (23,496 nút).

Bảng 5.1 *So sánh độ chính xác các CSDL tuần tự thu gọn bằng giải pháp PageRank tích hợp với CPT+*

TT	Kích cỡ (%)	Độ chính xác dự đoán truy cập Web của các cơ sở dữ liệu tuần tự thu gọn (%)										
		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Trung Bình
1	**100**	**99.927**	**99.941**	**99.936**	**99.924**	**99.938**	**99.938**	**99.941**	**99.934**	**99.936**	**99.947**	**99.936**
2	**98**	99.921	99.958	99.937	99.925	99.941	99.951	99.951	99.942	99.935	99.946	99.941
3	**96**	99.931	99.97	99.944	99.933	99.946	99.954	99.952	99.956	99.949	99.947	99.948
4	**94**	99.937	99.964	99.949	99.943	99.954	99.966	99.959	99.957	99.959	99.957	99.955
5	**92**	99.94	99.977	99.955	99.959	99.966	99.966	99.959	99.966	99.957	99.959	99.96
6	**90**	99.956	99.976	99.967	99.962	99.964	99.967	99.956	99.964	99.962	99.964	99.964
7	**88**	99.954	99.978	99.969	99.969	99.963	99.974	99.969	99.971	99.965	99.972	99.968
8	**86**	99.96	99.979	99.974	99.97	99.966	99.979	99.981	99.962	99.97	99.966	99.971
9	**84**	99.967	99.979	99.979	99.975	99.969	99.977	99.979	99.961	99.967	99.971	99.972
10	**82**	99.976	99.98	99.974	99.982	99.974	99.978	99.986	99.966	99.972	99.97	99.976
11	**80**	99.973	99.984	99.982	99.977	99.997	99.978	99.988	99.969	99.973	99.84	99.966
12	**78**	99.977	99.989	99.985	99.981	99.981	99.99	99.985	99.983	99.968	99.981	99.982
13	**76**	99.981	99.985	99.987	99.987	99.974	99.996	99.983	99.987	99.978	99.987	99.985
14	**74**	99.982	99.998	99.998	99.991	99.984	99.998	99.989	99.991	99.978	99.993	99.99
15	**72**	99.986	99.993	100	99.991	99.991	99.998	99.988	100	99.984	100	99.993
16	**70**	99.988	99.998	99.998	99.998	99.998	99.998	99.998	99.998	99.995	100	99.997
17	**68**	99.998	100	99.998	99.998	100	99.995	99.998	99.993	99.998	100	99.998
18	**66**	100	100	99.995	99.997	100	99.997	100	99.997	99.997	100	99.998
19	**64**	99.997	100	99.954	99.997	100	99.938	100	99.997	99.997	99.957	99.985
20	**62**	100	99.997	99.941	99.997	99.997	99.955	100	99.947	99.997	99.93	99.979
21	**60**	99.997	**99.866**	**99.748**	99.994	**99.824**	**99.751**	99.997	**99.718**	**99.675**	**99.76**	**99.862**
22	**58**	**99.636**	99.613	99.84	**99.581**	99.543	99.892	**99.693**	99.718	99.515	99.844	99.69
23	**56**	99.634	99.61	99.81	99.571	99.542	99.89	99.693	99.716	99.511	99.835	99.683

Theo kết quả thử nghiệm được minh họa như hình, độ chính xác dự đoán trung bình của các cơ sở dữ liệu huấn luyện ban đầu (có kích cỡ 90,000) là 99.936%, khi loại bỏ các chuỗi dữ liệu thừa để cơ sở dữ liệu thu gọn đạt đến kích cỡ là 66% (59,400 dòng) thì độ chính xác dự đoán trung bình là 100% (tăng 0.0621%). ***Hình 5.3*** minh họa biểu đồ so sánh trung bình độ chính xác dự đoán trên các tập dữ liệu thu gọn về kích cỡ mà không mất đi tính chính xác dự đoán bằng giải pháp PageRank (Chương 3).

Kích cỡ (%)	Độ chính xác trung bình mô hình CPT+
100	99.936
98	99.941
96	99.948
94	99.955
92	99.960
90	99.964
88	99.968
86	99.971
84	99.972
82	99.976
80	99.966
78	99.982
76	99.985
74	99.990
72	99.993
70	99.997
68	99.998
66	99.998
64	99.985
62	99.979
60	99.862
58	99.690
56	99.683



Hình 5.3 *Xây dựng các tập dữ liệu huấn luyện và kiểm thử dự đoán.*

Nhận xét rằng, khi giảm kích cỡ còn 66%, độ chính xác đạt đỉnh là 100% và bắt đầu một quá trình suy thoái về độ chính xác khi kích cỡ còn 62% trở xuống.

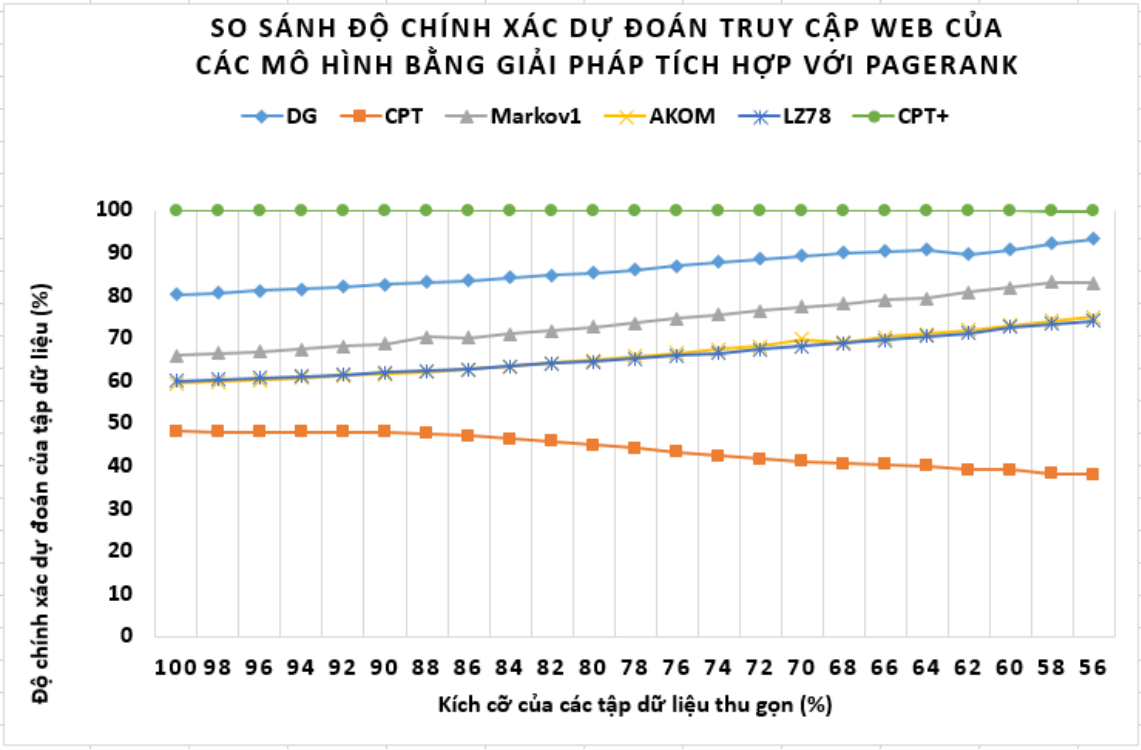
Từ kết quả thực nghiệm trên, ta có cơ sở để khẳng định rằng khi sử dụng tập dữ liệu huấn luyện thu gọn có kích cỡ 66 % (59,400) để tiếp tục cho giai đoạn tiếp là giai đoạn kiểm thử (dự đoán) là rất khả thi.

So sánh các mô hình dự đoán truy cập Web bằng cách tích hợp PageRank:

Kết quả thực nghiệm được trình chi tiết trong ***Bảng 5.2*** và ***Hình 5.4*** cho thấy rằng giải pháp tích hợp PageRank với CPT+ và DG là phù hợp với độ chính xác dự đoán truy cập Web là xấp xỉ đạt 100% đối với CPT+ và trên 80% đối với DG. Ngược lại giải pháp tích hợp PageRank với CPT (một phiên bản cũ của CPT+) là không phù hợp vì độ chính xác dự đoán truy cập Web chưa đạt đến 50%

Bảng 5.2 *Bảng thống kê độ chính xác của các mô hình tích hợp PageRank*

Kích cỡ (%)	Độ chính xác dự đoán truy cập Web trung bình					
	DG	CPT	Markov1	AKOM	LZ78	CPT+
100	80.116	48.088	65.932	59.451	59.945	99.936
98	80.585	48.031	66.338	59.773	60.319	99.941
96	81.060	48.007	66.799	60.171	60.630	99.948
94	81.486	47.946	67.312	60.591	60.922	99.955
92	81.996	47.955	67.986	61.145	61.415	99.960
90	82.499	47.924	68.580	61.577	61.811	99.964
88	83.044	47.631	70.282	62.123	62.267	99.968
86	83.517	46.994	70.076	62.702	62.707	99.971
84	84.087	46.332	70.925	63.353	63.265	99.972
82	84.678	45.741	71.728	64.083	64.002	99.976
80	85.292	44.877	72.592	64.829	64.502	99.966
78	85.931	44.091	73.495	65.780	65.095	99.982
76	86.828	43.295	74.501	66.478	65.897	99.985
74	87.834	42.397	75.466	67.283	66.486	99.990
72	88.497	41.700	76.362	68.146	67.260	99.993
70	89.311	41.105	77.278	69.733	68.037	99.997
68	89.931	40.611	78.080	68.887	68.773	99.998
66	90.307	40.281	78.918	70.293	69.443	99.998
64	90.781	39.932	79.299	70.986	70.407	99.985
62	89.613	39.124	80.771	71.879	71.195	99.979
60	90.731	39.084	81.868	72.873	72.534	99.862
58	92.086	38.125	83.127	73.959	73.317	99.690
56	93.304	37.970	82.887	74.931	73.958	99.683



Hình 5.4 *Biểu đồ so sánh độ chính xác dự đoán truy cập web của các mô hình bằng giải pháp tích hợp với PageRank*

Bên cạnh đó, ***Hình 5.4*** cũng cho thấy rằng khi tích hợp PageRank với CPT+ thì hiệu quả hơn tất cả các phương pháp còn lại (DG, Markov1, AKOM, LZ78, CPT). Do đó giải pháp tích hợp PageRank với CPT+ là giải pháp hiệu quả cho dự đoán truy cập Web.

5.2.3. Kết hợp giải pháp nâng cao độ chính xác và hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web

5.2.3.1. Mục đích

Chứng minh bằng thử nghiệm giải pháp tích hợp tính toán PageRank, phân tích chuỗi dữ liệu tuần tự, CPT+ đạt hiệu quả về thời gian dự đoán mà không làm mất đi độ chính xác dự đoán.

5.2.3.2. Dữ liệu

Dữ liệu được khai phá là 10 cơ sở dữ liệu tuần tự thu gọn có kích cỡ 66% so với cơ sở dữ liệu tuần tự huấn luyện gốc như đã được xây dựng ở phần trên. Mỗi cơ sở dữ liệu tuần tự này có số dòng là 54,900 và có độ chính xác dự đoán là 100% (điều này đã được chứng minh qua thử nghiệm ở phần trên).

5.2.3.3. Phương pháp

Nghiên cứu sinh tiến hành kiểm thử bằng cách dự đoán các chuỗi tuần tự con thuộc tập dự đoán (10% so với cơ sở dữ liệu huấn luyện gốc) trên 2 loại bộ dữ liệu huấn luyện khác nhau:

Bộ dữ liệu thứ nhất : 10 cơ sở dữ liệu tuần tự huấn luyện (90,000 dòng)

Bộ dữ liệu thứ hai: 10 cơ sở dữ liệu tuần tự huấn luyện thu gọn bằng kỹ thuật PageRank (54,900 dòng)

Trình tự thực hiện (10 lần trên từng tập cơ sở dữ liệu huấn luyện khác nhau):

Nhập vào một chuỗi tuần tự và cơ sở dữ liệu huấn luyện (90,000 dòng) áp dụng CPT+ để dự đoán và ghi nhận thời gian t_{90} (tính bằng milliseconds) và độ chính xác Acc_{90} của cơ sở dữ liệu huấn luyện này. Tiếp tục thực hiện dự đoán chuỗi tuần tự này trên cơ sở dữ liệu tuần tự thu gọn (54,000 dòng) bằng cách áp dụng CPT+ và kỹ thuật phân tích chuỗi (Chương 4) để thu được cơ sở dữ liệu tuần tự nhỏ nhỏ

hơn nhiều so với cơ sở dữ liệu tuần tự nhập vào và ghi nhận thời gian t_{66tiny} (tính bằng milliseconds) và độ chính xác Acc_{66tiny} của cơ sở dữ liệu thu gọn mới này.
So sánh t_{90} và t_{66tiny} để đưa ra kết luận về độ hiệu quả về thời gian dự đoán và so sánh Acc_{90} và Acc_{66tiny} và đưa ra kết luận việc thực hiện dự đoán như vật có mất đi tính chính xác hay không.

5.2.3.4. Các độ đo đánh giá

❖ Độ đo đánh giá về độ chính xác:

Áp dụng công thức *(1.1)* cho Acc_{90} và Acc_{66tiny}

Nếu $Acc_{66tiny} \geq Acc_{90}$: Dự đoán hiệu quả, ngược lại thì dự đoán không hiệu quả

Độ đo đánh giá về thời gian:

❖ Độ đo đánh giá về thời gian:

Nếu t_{66tiny} nhỏ hơn rất nhiều so với t_{90} : Dự đoán hiệu quả về thời gian, ngược lại thì dự đoán không hiệu quả về thời gian.

5.2.3.5. Kết quả thực nghiệm và phân tích

Nghiên cứu sinh tiến hành thực hiện dự đoán 200 chuỗi dữ liệu, Bảng 5.2 minh họa 10 chuỗi cần được dự đoán cùng với các thông tin về thời gian thực hiện dự đoán t_{90}, thời gian thực hiện dự đoán t_{66tiny} và kích cỡ của cơ sở dữ liệu được thu gọn nhờ vào kỹ thuật phân tích chuỗi mà được trình bày chi tiết ở Chương 4.

Bảng 5.3 *Minh họa hiệu quả về thời gian dự đoán*

TT	**Chuỗi dự đoán**	t_{90} *(milliseconds)*	T_{60tiny} *(milliseconds)*	**Kích cỡ của các cơ sở dữ liệu tuần tự được thu gọn** *(dòng)*
1	$\langle 6, 273, 77 \rangle$	30872	292	80
2	$\langle 40, 6, 90 \rangle$	26735	286	36
3	$\langle 6, 136, 1101 \rangle$	27628	294	116
4	$\langle 3, 64, 77 \rangle$	28471	316	149

5	⟨69, 3, 148⟩	29066	280	128
6	⟨215, 148, 303⟩	32373	262	16
7	⟨14, 64, 77⟩	31451	283	27
8	⟨27, 7, 87⟩	26941	344	502
9	⟨205, 69, 148⟩	35609	317	21
10	⟨303, 11, 7⟩	28577	258	95

Kiểm định ***Paired-Sample T Tests*** cho tập số liệu của t_{90} và t_{66tiny} được minh họa theo hình sau.

Phương pháp	**Thời gian thực thi trung bình (Mean)** *(milliseconds)*	**Giá trị thống kê (pValue)**
Dự đoán chỉ dùng phương pháp CPT+	28550.37	0.000 (< 0.01)
Dự đoán dùng giải pháp tích hợp CPT+, PageRank và phân tích chuỗi	354.91	

Kết quả thử nghiệm thu được trên hình chỉ ra rằng khi dự đoán chỉ dùng phương pháp CPT+ có rất chậm so với giải pháp tích hợp PageRank, CPT+ và phân tích chuỗi xấp xỉ 80 lần.

Thử nghiệm cũng cho thấy Acc_{66tiny} luôn trội hơn Acc_{90} cho dù là không đáng kể (xấp xỉ 0.0621%)

5.3. Kết luận Chương 5

Chương này đã trình bày đề xuất một giải pháp tổng hợp: Vừa nâng cao độ chính xác, vừa nâng cao hiệu năng về thời gian khai phá dữ liệu tuần tự cho dự

đoán truy cập Web. Kết quả thực nghiệm trên tập dữ liệu Kosarak (tập dữ liệu lớn nhất trong các nghiên cứu của luận án) cho thấy rằng khi kết hợp giải pháp ở Chương 3 và giải pháp ở Chương 4 thì có thể tăng độ chính xác trung bình lên 0.0621% và thời gian thực thi dự đoán trung bình hiệu quả hơn phương pháp truyền thống (chỉ áp dụng CPT+) lên đến 80 lần. Giải pháp cũng có một công trình liên quan là bài báo [CT9], [CT10].

PHẦN KẾT LUẬN

Phần kết luận tóm tắt lại ngắn gọn những đóng góp của luận án, bàn luận về các kết quả thu được, đồng thời nêu lên các mặt còn hạn chế và hướng phát triển của công trình nghiên cứu của luận án.

1. Đóng góp của luận án

Luận án trình bày 3 giải pháp cho dự đoán truy cập Web: (1) Giải pháp thiết kế và chuẩn hóa cơ sở dữ liệu tuần tự cho dự đoán truy cập Web; (2) Giải pháp nâng cao độ chính xác cho dự đoán truy cập Web; (3) Giải pháp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web.

2. Đánh giá, bàn luận tổng quan dự đoán truy cập Web

Công trình nghiên cứu CT5 trình bày tổng quan các tiếp cận khai phá dữ liệu để dự đoán truy cập Web. Trong công trình này, nghiên cứu sinh đã trình bày các phương pháp tiếp cận khác nhau để giải quyết bài toán dự đoán truy cập Web. Chẳng hạn nhiều nhà nghiên cứu đã dự đoán truy cập Web bằng các mô hình luật kết hợp, các mô hình dự đoán chuỗi tuần tự như Markov, DG, CPT,…các phương pháp lai. Tuy nhiên những mô hình này còn nhiều hạn chế là mất thông tin và thiếu chính xác và thời gian xử lý khá chậm. Trong các phương pháp giải quyết bài toán dự đoán truy cập Web, phương pháp CPT+ được đánh giá là trội hơn cả. Từ cơ sở này, nghiên cứu sinh đã đề xuất các giải pháp để dự đoán với CPT+ mang tính hiệu quả hơn bằng cách tích hợp các giải pháp nâng cao hiệu quả về độ chính xác và thời gian dự đoán. Chi tiết các giải pháp sẽ được bàn luận kỹ hơn trong các phần tiếp theo trong chương này.

2.1. Đánh giá, bàn luận về kết quả nghiên cứu chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web

Chương 2 trình bày nghiên cứu xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web. Các công trình liên quan đến **Chương 2** bao gồm các công trình [CT2], [CT3], [CT6].

Cụ thể, [CT2] trình bày tiếp cận khai phá dữ liệu tuần tự với khai luật liên tiếp dựa trên ER-Miner . Mặc dù ER-Miner có ưu điểm hơn các nghiên cứu khai phá dữ liệu để dự đoán truy cập Web khác như các nghiên cứu dựa trên Sequential Pattern Mining, Sequential Rules Mining (CMDeo [32], CMRules [32], RuleGrowth [39]). Các nghiên cứu dựa trên Sequential Pattern Mining [4] bao gồm việc tìm kiếm các chuỗi con xuất hiện một cách thường xuyên trong một tập các chuỗi dữ liệu tuần tự. Tuy nhiên, một chuỗi tuần tự xuất hiện một cách thường xuyên thì không đủ để thực hiện dự đoán. Các mô hình tốt hơn là sử dụng Sequential Rules Mining như các mô hình CMDeo, CMRules [6]. CMDeo là một giải thuật khai phá không gian tìm kiếm của luật dùng phương pháp tìm kiếm theo chiều rộng. Cách này có hạn chế là tạo ra quá nhiều ứng viên. Một phương pháp thay thế là CMRules được đề xuất, phương pháp này dựa vào tính chất là bất kỳ luật liên tiếp nào cũng phải là luật kết hợp để cắt tỉa không gian tìm kiếm của các luật liên tiếp [37]. Nó cho thấy thực thi nhanh hơn CMDeo. Tiếp đó, RuleGrowth được đề xuất. Đây là kỹ thuật dựa vào tiếp cận phát triển mẫu để tránh tạo ra ứng viên. Tuy nhiên, đối với các tập dữ liệu dày đặt và các chuỗi tuần tự dài, phương pháp này không hiệu quả vì tốn chi phí về thời gian vì thực hiện nhiều phép chiếu trên cơ sở dữ liệu [46]. Tuy nhiên, khai phá luật liên tiếp với ER-Miner vẫn còn nhiều hạn chế so với các nghiên cứu dùng tiếp cận dùng dự đoán chuỗi tuần tự (sequence prediction). Các mô hình khai phá luật liên tiếp (điển hình là ER-Miner) cũng như mô hình dựa trên Markov đều được xây dựng bằng cách dùng một phần thông tin chứa trong các chuỗi tuần tự huấn luyện để thực hiện dự đoán do vậy có

một sự giảm độ chính xác đáng kể. Do vậy, một giải pháp thay thế là áp dụng CPT+. Mô hình CPT+ là một phiên bản cải tiến so với CPT mà dựa vào cấu trúc cây dự đoán nén để nâng cao độ chính xác trong dự đoán dữ liệu tuần tự. Điểm nổi bật của CPT+ là giải pháp này cung cấp 3 chiến lược hiệu quả để làm giảm kích cỡ của cây dự đoán nén, giảm thời gian dự đoán và nâng cao độ chính xác cho CPT.

[CT3] mô tả việc xây dựng cơ sở dữ liệu tuần tự từ dữ liệu nhật ký Web để phục vụ cho dự đoán truy cập Web. CT3 đã chỉ ra những hạn chế khi dự đoán truy cập Web mà chưa biến đổi dữ liệu nhật ký Web thành cơ sở dữ liệu và đưa ra giải pháp để thiết kế cơ sở dữ liệu tuần tự để dự đoán truy cập Web. Theo đó, CT3 cũng đã chứng minh rằng cơ sở dữ liệu tuần tự được xây dựng là phù hợp cho dự đoán truy cập Web bằng CPT+ với độ chính xác cao hơn các tiếp cận phổ biến còn lại như CPT, DG, 1st Markov, All-k-Markov…

Bên cạnh đó, [CT6] đề xuất một giải pháp dự đoán truy cập Web cho Website bán hàng, cụ thể [CT6] trình bày giải pháp biến đổi cơ sở dữ liệu đặt hàng (một phần của cơ sở dữ liệu quan hệ) sang cơ sở dữ liệu tuần tự để dự đoán truy cập Web. Ngoài ra phương pháp biến đổi cơ sở dữ liệu quan hệ, cơ sở dữ liệu mạng có nhãn thời gian cũng được nghiên cứu sinh thực hiện trong công trình [CT8].

Để minh họa rõ hơn giải pháp chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web, ***Bảng 6.1*** trình bày so sánh giải pháp chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web theo kỹ thuật tuần tự và song song.

Gọi T_{tt} là thời gian thực thi giải pháp chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web theo kỹ thuật xử lý tuần tự và T_{ss} là thời gian thực thi giải pháp chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web theo kỹ thuật xử lý song song.

Ta có $T_{tt} = \theta T_{ss}$, với θ_1 là hằng số cho biết rằng T_{ss} nhanh gấp θ_1 lần T_{tt}.

Bảng 6.1 *So sánh giải pháp chuẩn hóa cơ sở dữ liệu Web Log cho dự đoán truy cập Web theo kỹ thuật tuần tự và song song*

Bộ dữ liệu	T_{tt} (milliseconds)	T_{ss} (milliseconds)	θ_1 (lần)
periwinkletecottages.com	121836	97449	1.25
palmviewsanibel.com	85683	74814	1.15
devqa.robotec.co.il	12382	9893	1.25
inees.org	3508	3312	1.06

2.2. Đánh giá, bàn luận về kết quả nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web

Từ kết quả nghiên cứu ở **Chương 2**, nghiên cứu sinh đã đề xuất giải pháp nâng cao hiệu quả về độ chính xác khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Cụ thể, Chương 4 trình bày kết quả nghiên cứu của [CT7], nghiên cứu sinh đã chứng minh rằng giải pháp dự đoán truy cập Web tích hợp CPT+ với giải thuật PageRank đã có hiệu quả tốt hơn so với phương pháp chỉ sử dụng CPT+ về độ chính xác dự đoán.

Với các tập dữ liệu MSNBC, FIFA, KOSARAK, nghiên cứu sinh đã thực hiện giảm đến 50%, 15%, 30% (theo trình tự các tập dữ liệu) kích cỡ không gian dự đoán (kích cỡ của cơ sở dữ liệu tuần tự) nhưng độ chính xác của giải pháp tích hợp giải thuật PageRank với CPT+ vẫn luôn cao hơn độ chính xác của tiếp cận chỉ dùng CPT+ (kích cỡ cơ sở dữ liệu tuần tự chưa giảm kích cỡ).

Để minh họa rõ hơn giải pháp nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web, ***Bảng 6.2*** trình bày kết quả so sánh giải pháp nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web.

Gọi Max_reduction (%) là độ giảm kích cỡ tối đa so với kích cỡ cơ sở dữ liệu tuần tự gốc và Acc_Reduction (%) là độ chính xác của cơ sở dữ liệu tuần tự đã loại

bỏ Max_reduction (%) về kích cỡ của cơ sở dữ liệu tuần tự gốc, và Acc_Origin là độ chính xác của cơ sở dữ liệu gốc, *Bảng 6.2* sẽ trình bày chi tiết các giá trị này.

Bảng 6.2 *So sánh giải pháp nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web*

Bộ dữ liệu	Max_Reduction (%)	Acc_ Reduction (%)	Acc_Origin %
MSNBC	40	71.616	46.389
FIFA	16	99.899	99.888
KOSARAK	32	99.951	99.947

2.3. Đánh giá, bàn luận về kết quả nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web

Từ kết quả nghiên cứu ở **Chương 2**, nghiên cứu sinh đã đề xuất giải pháp nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web. Cụ thể, Chương 3 trình bày kết quả nghiên cứu của [CT1] và [CT4]. Trong các công trình nghiên cứu này, nghiên cứu sinh đã thử nghiệm giải pháp dự đoán chuỗi tuần tự cải tiến bằng cách tích hợp phương pháp phân tích chuỗi dự đoán với phương pháp CPT+, cụ thể là việc dự đoán phụ thuộc vào chuỗi tuần tự cần dự đoán. Bằng cách loại bỏ chuỗi tuần tự dư thừa, không có ý nghĩa, điều này đồng thời cũng làm giảm kích cỡ không gian dự đoán của cơ sở dữ liệu tuần để dự đoán được hiệu quả hơn về thời gian. Nghiên cứu sinh đã thử nghiệm trên 3 tập dữ liệu lớn là FIFA, MSBNC, KOSARAK và thu được các kết quả hiệu năng thời gian tốt hơn rất nhiều lần (nhanh hơn về tốc độ thực thi dự đoán) nhưng vẫn không mất đi tính chính xác của dự đoán. Để minh họa rõ hơn giải pháp nâng cao hiệu quả về thời gian dự đoán truy cập Web, ***Bảng 5.3*** trình bày kết quả so sánh giải pháp nâng cao hiệu quả về thời gian thực thi dự đoán truy cập Web.

Gọi T_I là thời gian thực thi dự đoán chuỗi theo phương pháp chỉ dùng CPT+ và T_{II} là thời gian thực thi dự đoán chuỗi theo giải pháp kết hợp xử lý chuỗi và CPT+, trong đó θ_2 là giá trị cho biết T_{II} có thời gian thực thi nhanh gấp θ_2 lần T_I. *Bảng 6.3* sẽ trình bày chi tiết các giá trị này.

Bảng 6.3 *So sánh giải pháp nâng cao hiệu quả về thời gian thực thi dự đoán truy cập Web*

Bộ dữ liệu	T_I (milliseconds)	T_{II} (milliseconds)	θ_2 (lần)
FIFA	18763.55	6092.15	3.08
KOSARAK	28166.08	912.17	30.88
BMS	36361.56	351.39	103.50

2.4. Đánh giá, bàn luận về kết quả kết hợp giải pháp nâng cao độ chính xác và nâng cao hiệu quả về thời gian khai phá dữ liệu tuần tự cho dự đoán truy cập Web

Xét 200 mẫu dự đoán thực hiện trên các tập huấn luyện của cơ sở dữ liệu tuần tự Kosarak, *Bảng 6.4* trình bày tổng hợp các giải pháp cho dự đoán truy cập Web với thời gian thực thi trung bình và độ chính xác dự đoán trung bình khai phá dữ liệu cho dự đoán truy cập Web trên tập Kosarak.

Bảng 6.4 *Bảng tổng hợp thời gian thực thi trung bình và độ chính xác trung bình của các giải pháp cho dự đoán truy cập Web*

Các giải pháp	Thời gian thực thi trung bình (milliseconds)	Độ chính xác dự đoán trung bình (%)
Giải pháp truyền thống (dùng CPT+)	28550.37	99.39%
Giải pháp nâng cao độ chính xác và hiệu quả về thời gian (tích hợp PageRank, phân tích chuỗi và CPT+)	354.91	100%

Bảng 6.4 cho thấy kết hợp giải pháp nâng cao độ chính xác và nâng cao hiệu quả về thời gian khai phá dữ liệu cho dự đoán truy cập Web rất hiệu quả về mặt thời gian cũng như độ chính xác dự đoán.

2.5. Kết luận và kiến nghị

2.5.1 Ưu điểm

Qua quá trình thực hiện luận án, nghiên cứu sinh đã học hỏi được rất nhiều kiến thức liên quan đến xử lý dữ liệu Web Log, các mô hình dự đoán truy cập Web, những ưu điểm, những hạn chế của các mô hình này, đặc biệt là mô hình dự đoán chuỗi dữ liệu tuần tự cây dự đoán nén cải tiến (CPT+). Bên cạnh đó, kiến thức về giải thuật PageRank cũng rất hữu ích trong việc dự đoán truy cập Web dựa trên mối quan hệ giữa các liên kết.

Từ việc nghiên cứu tổng quan các phương pháp, cũng như các mô hình cho dự đoán hành vi truy cập Web, nghiên cứu sinh đã đề xuất các giải pháp khác nhau để giải quyết bài toán dự đoán truy cập Web như chuẩn hóa và xây dựng cơ sở dữ liệu tuần tự, cải tiến về thời gian và độ chính xác cho dự đoán truy cập Web với CPT+.

Bên cạnh đó, các công trình nghiên cứu liên quan đến luận án cũng đã được thực hiện và được đăng trên các Hội thảo, Tạp chí chuyên ngành trong nước và quốc tế. Cụ thể là, có 2 công trình thuộc Hội thảo trong nước ([CT1], [CT6]), 1 công trình thuộc Tạp chí trong nước ([CT2]), 2 công trình thuộc Hội thảo quốc tế ([CT5], [CT8]), 3 công trình thuộc Tạp chí quốc tế ([CT3]-ESCI, [CT4], [CT7]-Scopus, [CT9], [CT10] (đã được chấp nhận, chuẩn bị xuất bản)).

2.5.2 Hạn chế

Như đã trình bày ở trên, thời gian thực thi dự đoán của giải pháp được đề xuất (Tích hợp giải thuật, giải thuật phân tích chuỗi dự đoán và giải thuật CPT+) nhanh hơn rất nhiều lần so với thời thực thi theo phương pháp thông thường (chỉ dùng giải thuật CPT+). Tuy nhiên, để tăng độ chính xác cho dự đoán, quá trình tiền xử lý (cụ thể là tính toán PageRank của từng trang, tính toán PageRank cho từng chuỗi dữ liệu tuần tự) để loại bỏ các chuỗi dữ liệu dư thừa, không có ý nghĩa cho dự đoán tốn nhiều thời gian trong quá trình huấn luyện.

2.5.3. Hướng phát triển

Kết quả luận án mới chỉ là bước đầu trong quá trình nghiên cứu của nghiên cứu sinh, còn nhiều vấn đề về lý thuyết và áp dụng trong thực tiễn cần phải hoàn thiện hơn. Trong tương lai, nghiên cứu sinh đặc biệt quan tâm đến việc nâng cao kỹ thuật tính toán để có được những kết quả thực nghiệm tốt hơn. Sau đây là một số kế hoạch phát triển kết quả luận án trong tương lai:

+ Khai phá dữ liệu truy cập Web trên các tập dữ liệu click-stream trong các cơ sở dữ liệu rất lớn, Big Data để đánh giá hiệu quả của giải pháp được trình bày trong luận án.

+ Nghiên cứu thêm giải pháp tối ưu để khai phá dữ liệu cho dự đoán truy cập Web.

+ Áp dụng kết quả nghiên cứu của luận án để dự đoán truy cập Web của người học trong hệ thống E-Learning phục vụ cho đào tạo trực tuyến. Đặc biệt, nghiên cứu

sinh và các đồng sự đang viết một bài báo về mô hình dự báo xu hướng tăng giảm của các đồng tiền điện tử dựa trên các kết quả nghiên cứu đã thực hiện.

DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU

Luận án này là kết quả của những công trình nghiên cứu sau:

CT1. **Nguyễn Thôn Dã**, Tân Hạnh (12/2017). Một Giải Pháp Nâng Cao Hiệu Quả Cho Dự Đoán Chuỗi Dữ Liệu Tuần Tự. Hội thảo Quốc gia lần thứ XX về Điện tử, Truyền thông và Công nghệ Thông tin (National Conference on Electronics, Communications and Information Technology – REV-ECIT), TP.HCM. Công trình này (CT1) liên quan đến mục tiêu thứ tư của luận án.

CT2. **Nguyen Thon Da**, Tan Hanh (Dec-2017). Improving Performance of Sequential Rule Mining With Parallel Computing. Tạp chí Khoa học Công nghệ Thông tin và Truyền thông (JSTIC), Số 02&03. Trang 86-86, ISSN: 2525-2224. Công trình này (CT2) liên quan đến mục tiêu thứ nhất của luận án.

CT3. **Nguyen Thon Da**, Tan Hanh, Pham Hoang Duy (Feb-2018). An Approach To Build Sequence Database From Web Log Data For Webpage Access Prediction. International Journal of Computer Science and Network Security (IJCSNS), Vol. 18 No. 2 pp. 138-143, ISSN: 1738-7906. (ESCI). Công trình này (CT3) liên quan đến mục tiêu thứ hai của luận án.

CT4. **Nguyen Thon Da**, Tan Hanh (Sep-2018), A novel approach based on sequence prediction for webpage access, International Journal of Engineering & Technology, 7 (4) (2018) 2356-2359 (DOI: 10.14419/ijet.v7i4.13901). Công trình này (CT4) liên quan đến mục tiêu thứ tư của luận án.

CT5. **N. T. Da**, T. Hanh and P. H. Duy (2018), "A Survey of Webpage Access Prediction," 2018 International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, Vietnam, 2018, pp. 315-320. doi: 10.1109/ATC.2018.8587490 (ATC 2018). Công trình này (CT5) liên quan đến mục tiêu thứ nhất của luận án.

CT6. **Nguyễn Thôn Dã**, Tân Hạnh, Hồ Trung Thành (12-2018), Dự đoán hành vi đặt hàng dựa trên mô hình dự đoán chuỗi tuần tự, Hội thảo khoa học Hệ thống thông tin trong kinh doanh và quản lý (ISBM18), NXB ĐHQG. TPHCM, trang 260 - 274, ISBN 978-604-73-6504. Công trình này (CT6) liên quan đến mục tiêu thứ hai của luận án.

CT7. **Da, N. T.**, Hanh, T., & Duy, P. H. (2019). Improving webpage access predictions based on sequence prediction and pagerank algorithm. Interdisciplinary Journal of Information, Knowledge, and Management, Volume 14, p27-p44. https://doi.org/10.28945/4176 (Scopus, Q3). Công trình này (CT7) liên quan đến mục tiêu thứ ba của luận án.

CT8. **Da Nguyen Thon**, Hanh Tan and Duy Pham Hoang (2019), Sequence Prediction In Temporal Networks, 15th International Conference on Multimedia Information Technology and Application, ISSN: 1975-4736. Công trình này (CT8) liên quan đến mục tiêu thứ hai của luận án.

CT9. **Nguyen Thon Da**, Tan Hanh, Pham Hoang Duy (2020). Improving webpage access predictions based on sequence prediction and pagerank algorithm. International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-6, March 2020, p2327-p2335.

CT10. **Nguyen Thon Da**, Tan Hanh (2020). Investigating the PageRank and sequence prediction based approaches for next page prediction. International Journal of Electrical and Computer Engineering(IJECE), ISSN: 2088-8708 (Scopus, Q2) (Đã được chấp nhận chuẩn bị xuất bản). Công trình này (CT9) liên quan đến mục tiêu thứ ba của luận án.

TÀI LIỆU THAM KHẢO

[1] Abdulwahhab, R. S., & Abdulwahab, S. S. (2017). *Integrating learning analytics to predict student performance behavior.* Paper presented at the Information and Communication Technology and Accessibility (ICTA), 2017 6th International Conference on.

[2] Abraham, A. (2003). Business intelligence from web usage mining. *Journal of Information & Knowledge Management, 2*(04), 375-390.

[3] Adami, G., Avesani, P., & Sona, D. (2003). *Clustering documents in a web directory.* Paper presented at the Proceedings of the 5th ACM international workshop on Web information and data management.

[4] Agrawal, R., & Srikant, R. (1995). *Mining sequential patterns.* Paper presented at the icde.

[5] Amento, B., Terveen, L., & Hill, W. (2000). *Does "authority" mean quality? Predicting expert quality ratings of Web documents.* Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

[6] Anandhi, D., & Ahmed, M. I. (2017). Prediction of user's type and navigation pattern using clustering and classification algorithms. *Cluster Computing*, 1-10.

[7] Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications, 8*(11), 7-10.

[8] Awad, M., Khan, L., & Thuraisingham, B. (2008). Predicting WWW surfing using multiple evidence combination. *The VLDB Journal—The International Journal on Very Large Data Bases, 17*(3), 401-417.

[9] Awad, M. A., & Khalil, I. (2012). Prediction of user's web-browsing behavior: Application of markov model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(4), 1131-1142.

[10] Bahram, S., Sen, D., & Amant, R. S. (2011). *Prediction of web page accessibility based on structural and textual features.* Paper presented at the Proceedings of the International Cross-Disciplinary Conference on Web Accessibility.

[11] Beggs, C. B., Shepherd, S. J., Emmonds, S., & Jones, B. (2017). A novel application of PageRank and user preference algorithms for assessing the relative performance of track athletes in competition. *PloS one, 12*(6), e0178458.

[12] Bestavros, A. (1995). *Using speculation to reduce server load and service time on the WWW.* Paper presented at the Proceedings of the fourth international conference on Information and knowledge management.

[13] Bhargav, A., & Bhargav, M. (2014). *Pattern discovery and users classification through web usage mining.* Paper presented at the Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on.

[14] Bollen, J., Rodriquez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics, 69*(3), 669-687.

[15] Bonino, D., Corno, F., & Squillero, G. (2003). *A real-time evolutionary algorithm for web prediction.* Paper presented at the Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on.

[16] Bouras, C., Konidaris, A., & Kostoulas, D. (2004). Predictive prefetching on the web and its potential impact in the wide area. *World Wide Web, 7*(2), 143-179.

[17] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30*(1-7), 107-117.

[18] Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data mining and knowledge discovery, 7*(4), 399-424.

[19] Castellano, G., Fanelli, A. M., & Torsello, M. A. (2013). Web usage mining: Discovering usage patterns for web applications. In *Advanced Techniques in Web Intelligence-2* (pp. 75-104): Springer.

[20] Chembath, J., & Fredrik, E. T. (2017). An Empirical Analysis of Algorithms to Predict Next Web Page Using Web Log Data. *International Journal of Applied Engineering Research, 12*(16), 5648-5654.

[21] Chimphlee, S., Salim, N., Bin Ngadiman, M. S., & Chimphlee, W. (2006). Using association rules and markov model for predit next access on web usage mining. *Advances in Systems, Computing Sciences and Software Engineering*, 371-376.

[22] Chimphlee, S., Salim, N., Ngadiman, M. S. B., & Chimphlee, W. (2006). Using association rules and markov model for predit next access on web usage mining. In *Advances in Systems, Computing Sciences and Software Engineering* (pp. 371-376): Springer.

[23] Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE transactions on Communications, 32*(4), 396-402.

[24] da Costa, M. G., & Gong, Z. (2005). *Web structure mining: an introduction.* Paper presented at the Information Acquisition, 2005 IEEE International Conference on.

[25] Dhyani, D., Bhowmick, S., & Ng, W.-K. (2003). *Modelling and predicting a Web page accesses using Markov processes.* Paper presented at the Database and Expert Systems Applications, 2003. Proceedings. 14th International Workshop on.

[26] Dongshan, X., & Junyi, S. (2002). A new markov model for web access prediction. *Computing in Science & Engineering, 4*(6), 34-39.

[27] Dubey, S., & Mishra, N. (2011). Web page prediction using hybrid model. *International Journal on Computer Science and Engineering, 3*(5), 2170-2176.

[28] Dutta, R., Kundu, A., Dattagupta, R., & Mukhopadhyay, D. (2009). An approach to web page prediction using markov model and web PageRanking. *Journal of Convergence Information Technology, 4*(4), 61-67.

[29] Dutta, R., Kundu, A., & Mukhopadhyay, D. (2011). Clustering-based web page prediction. *International Journal of Knowledge and Web Intelligence, 2*(4), 257-271.

[30] Eichinger, F., Nauck, D. D., & Klawonn, F. (2006). *Sequence mining for customer behaviour predictions in telecommunications.* Paper presented at the Proceedings of the Workshop on Practical Data Mining at ECML/PKDD.

[31] Eirinaki, M., Vazirgiannis, M., & Kapogiannis, D. (2005). *Web path recommendations based on PageRanking and markov models.* Paper presented at the Proceedings of the 7th annual ACM international workshop on Web information and data management.

[32] Fournier-Viger, P., Faghihi, U., Nkambou, R., & Nguifo, E. M. (2012). CMRules: Mining sequential rules common to several sequences. *Knowledge-Based Systems, 25*(1), 63-76.

[33] Fournier-Viger, P., Gomariz, A., Campos, M., & Thomas, R. (2014). *Fast vertical mining of sequential patterns using co-occurrence information.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[34] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., & Tseng, V. S. (2014). SPMF: a Java open-source pattern mining library. *The Journal of Machine Learning Research, 15*(1), 3389-3393.

[35] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., & Tseng, V. S. (2014). SPMF: A Java Open-source Pattern Mining Library. *Journal of Machine Learning Research, 15*(1), 3389-3393.

[36] Fournier-Viger, P., Gueniche, T., & Tseng, V. S. (2012). *Using Partially-Ordered Sequential Rules to Generate More Accurate Sequence Prediction.* Paper presented at the ADMA.

[37] Fournier-Viger, P., Gueniche, T., Zida, S., & Tseng, V. S. (2014). *ERMiner: sequential rule mining using equivalence classes.* Paper presented at the International Symposium on Intelligent Data Analysis.

[38] Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition, 1*(1), 54-77.

[39] Fournier-Viger, P., Nkambou, R., & Tseng, V. S.-M. (2011). *RuleGrowth: mining sequential rules common to several sequences by pattern-growth.* Paper presented at the Proceedings of the 2011 ACM symposium on applied computing.

[40] Frias-Martinez, E., & Karamcheti, V. (2002). *A prediction model for user access sequences.* Paper presented at the WEBKDD Workshop: Web Mining for Usage Patterns and User Profiles.

[41] García, E., Romero, C., Ventura, S., & Calders, T. (2007). *Drawbacks and solutions of applying association rule mining in learning management systems.* Paper presented at the Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007), Crete, Greece.

[42] García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*: Springer.

[43] Geetharamani, R., Revathy, P., & Jacob, S. G. (2015). Prediction of users webpage access behaviour using association rule mining. *Sadhana, 40*(8), 2353-2365.

[44] Géry, M., & Haddad, H. (2003). *Evaluation of web usage mining approaches for user's next request prediction.* Paper presented at the Proceedings of the 5th ACM international workshop on Web information and data management.

[45] Gopalakrishnan, T., Sengottuvelan, P., Bharathi, A., & Lokeshkumar, R. (2018). An Approach To Webpage Prediction Method Using Variable Order Markov

Model In Recommendation Systems. *Journal of Internet Technology, 19*(2), 415-424.

[46] Gueniche, T., Fournier-Viger, P., Raman, R., & Tseng, V. S. (2015). *CPT+: Decreasing the time/space complexity of the Compact Prediction Tree.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[47] Gueniche, T., Fournier-Viger, P., & Tseng, V. S. (2013). *Compact Prediction Tree: A Lossless Model for Accurate Sequence Prediction.* Paper presented at the ADMA (2).

[48] Gueniche, T., Fournier-Viger, P., & Tseng, V. S. (2013). *Compact prediction tree: A lossless model for accurate sequence prediction.* Paper presented at the International Conference on Advanced Data Mining and Applications.

[49] Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M., et al. (2013). Effective web log mining and online navigational pattern prediction. *Knowledge-Based Systems, 49*, 50-62.

[50] Gündüz, Ş., & Özsu, M. T. (2003). *A web page prediction model based on click-stream tree representation of user behavior.* Paper presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.

[51] Gupta, S., & Singhal, A. (2017). *Phishing URL detection by using artificial neural network with PSO.* Paper presented at the 2017 2nd International Conference on Telecommunication and Networks (TEL-NET).

[52] Hassan, M. T., Junejo, K. N., & Karim, A. (2009). *Learning and predicting key Web navigation patterns using Bayesian models.* Paper presented at the International Conference on Computational Science and Its Applications.

[53] Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*: MIT press.

[54] Ho, J., Lukov, L., & Chawla, S. (2005). *Sequential pattern mining with constraints on large protein databases.* Paper presented at the Proceedings of the 12th International Conference on Management of Data (COMAD).

[55] Hoekstra, J. (2016). *Predicting train journeys from smart card data: a real-world application of the sequence prediction problem.*

[56] Hornik, K., Grün, B., & Hahsler, M. (2005). arules-A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software, 14*(15), 1-25.

[57] Iliopoulos, C. S., Makris, C., Panagis, Y., Perdikuri, K., Theodoridis, E., & Tsakalidis, A. (2006). The weighted suffix tree: an efficient data structure for handling molecular weighted sequences and its applications. *Fundamenta Informaticae, 71*(2, 3), 259-277.

[58] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

[59] Jespersen, S., Pedersen, T. B., & Thorhauge, J. (2003). *Evaluating the markov assumption for web usage mining.* Paper presented at the Proceedings of the 5th ACM international workshop on Web information and data management.

[60] Jianhui, L., & Bingjie, Z. (2009). *A Web Prediction Pattern Recommendation Algorithm.* Paper presented at the Networking and Digital Society, 2009. ICNDS'09. International Conference on.

[61] Khalil, F., Li, J., & Wang, H. (2006). *A framework of combining Markov model with association rules for predicting web page accesses.* Paper presented at the Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61.

[62] Khalil, F., Li, J., & Wang, H. (2008). *Integrating recommendation models for improved web page prediction accuracy.* Paper presented at the Proceedings of the thirty-first Australasian conference on Computer science-Volume 74.

[63] Khalil, F., Li, J., & Wang, H. (2009). An integrated model for next page access prediction. *International Journal of Knowledge and Web Intelligence, 1*(1-2), 48-80.

[64] Khalil, F., Li, J., & Wang, H. (2009). An integrated model for next page access prediction. *IJ Knowledge and Web Intelligence, 1*(1/2), 48-80.

[65] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604-632.

[66] Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.

[67] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.

[68] Kumar, B. H., Vibha, L., & Venugopal, K. (2016). *Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order Markov model.* Paper presented at the Region 10 Symposium (TENSYMP), 2016 IEEE.

[69] Kumar, P., Kadambari, S., & Rawat, S. (2015). *Prefetching web pages for Improving user access latency using integrated Web Usage Mining.* Paper presented at the Communication, Control and Intelligent Systems (CCIS), 2015.

[70] Kundra, K., Kaur, U., & Singh, D. (2015). Efficient Web Log Mining and Navigational Prediction with EHPSO and Scaled Markov Model. In *Computational Intelligence in Data Mining-Volume 3* (pp. 529-543): Springer.

[71] Labroche, N., Monmarché, N., & Venturini, G. (2002). *A new clustering algorithm based on the chemical recognition system of ants.* Paper presented at the Proceedings of the 15th European Conference on Artificial Intelligence.

[72] Laird, P., & Saul, R. (1994). Discrete sequence prediction and its applications. *Machine learning, 15*(1), 43-68.

[73] Li, J.-Q., Zhao, Y., & Garcia-Molina, H. (2012). A path-based approach for web page retrieval. *World Wide Web, 15*(3), 257-283.

[74] Li, M., Yu, X., & Ryu, K. H. (2014). MapReduce-based web mining for prediction of web-user navigation. *Journal of Information Science, 40*(5), 557-567.

[75] Lin, W.-Y., Tseng, M.-C., & Su, J.-H. (2002). *A confidence-lift support specification for interesting associations mining.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[76] Liraki, Z., Harounabadi, A., & Mirabedini, J. (2015). Predicting the Users' Navigation Patterns in Web, using Weighted Association Rules and Users'

Navigation Information. *International Journal of Computer Applications, 110*(12).

[77] Luotonen, A. (1995). The common log file format.

[78] Maurya, J., Singh, S., Patil, H., & Jain, P. (2014). A Survey on: Methods of Web Behavior Prediction by: Utilizing Different Features. *International Journal, 4*(3).

[79] Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). *Using sequential and non-sequential patterns in predictive web usage mining tasks.* Paper presented at the Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on.

[80] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications, 25*(2), 443-458.

[81] Morse, P. M. (1968). Library effectiveness: A systems approach.

[82] Narvekar, M., & Banu, S. S. (2015). Predicting user's Web navigation behavior using hybrid approach. *Procedia Computer Science, 45*, 3-12.

[83] Nigam, B., Tokekar, S., & Jain, S. (2015). Evaluation of models for predicting user's next request in web usage mining. *international Journal on Cybernetics & informatics (UCi), 4*, 1-13.

[84] Padmanabhan, V. N., & Mogul, J. C. (1996). Using predictive prefetching to improve world wide web latency. *ACM SIGCOMM Computer Communication Review, 26*(3), 22-36.

[85] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*: Stanford InfoLab.

[86] Papadakis, N. K., Skoutas, D., Raftopoulos, K., & Varvarigou, T. A. (2005). Stavies: A system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques. *IEEE Transactions on Knowledge and Data Engineering, 17*(12), 1638-1652.

[87] Papapetrou, P., Kollios, G., Sclaroff, S., & Gunopulos, D. (2005). *Discovering frequent arrangements of temporal intervals.* Paper presented at the null.

[88] Patil, N. V., & Patil, H. D. Prediction of Web User's Browsing Behavior using All Kth Markov model and CSB-mine.

[89] Pei, J., Han, J., Mortazavi-Asl, B., & Zhu, H. (2000). *Mining access patterns efficiently from web logs.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[90] Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User modeling and user-adapted interaction, 13*(4), 311-372.

[91] Pirolli, P. L., & Pitkow, J. E. (1999). Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web, 2*(1-2), 29-45.

[92] Pitkänen, H. (2017). Exploratory sequential data analysis of user interaction in contemporary BIM applications.

[93] Pitkow, J., & Pirolli, P. (1999). *Mininglongestrepeatin g subsequencestopredict worldwidewebsurfing.* Paper presented at the Proc. UsENIX symp. on Internet Technologies and systems.

[94] Poornalatha, G., Chetan, S., & Raghavendra, P. S. (2017). Prediction model for prefetching web page ased on the usage patter. *International Journal of Control Theory and Applications, 10*(14), 39-47.

[95] Rao, V. M., & Kumari, V. V. (2010). An efficient hybrid successive Markov model for predicting web user usage behavior using web usage mining. *International Journal of Data Engineering (IJDE), 1*(5), 43-62.

[96] Rathod, V. R., & Patel, G. V. (2016). Prediction of User Behavior using Web log in Web Usage Mining. *International Journal of Computer Applications, 139*(8).

[97] Rigou, M., Sirmakessis, S., & Tzimas, G. (2006). *A method for personalized clustering in data intensive web applications.* Paper presented at the Proceedings of the joint international workshop on Adaptivity, personalization & the semantic web.

[98] Rjeily, C. B., Badr, G., Al Hassani, A. H., & Andres, E. (2017). *Predicting heart failure class using a sequence prediction algorithm.* Paper presented at the Advances in Biomedical Engineering (ICABME), 2017 Fourth International Conference on.

[99] Rjeily, C. B., Badr, G., El Hassani, A. H., & Andres, E. (2019). Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field. In *Machine Learning Paradigms* (pp. 71-99): Springer.

[100] Sampath, P., Wahi, A., & Ramya, D. (2014). A COMPARATIVE ANALYSIS OF MARKOV MODEL WITH CLUSTERING AND ASSOCIATION RULE MINING FOR BETTER WEB PAGE PREDICTION. *Journal of Theoretical & Applied Information Technology, 63*(3).

[101] Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. *Computer Networks, 33*(1), 377-386.

[102] Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains1. *Computer Networks, 33*(1-6), 377-386.

[103] Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Www, 1*, 285-295.

[104] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter, 1*(2), 12-23.

[105] Srivastava, T., Desikan, P., & Kumar, V. (2005). Web mining–concepts, applications and research directions. In *Foundations and advances in data mining* (pp. 275-307): Springer.

[106] Strehl, A., Ghosh, J., & Mooney, R. (2000). *Impact of similarity measures on web-page clustering.* Paper presented at the Workshop on artificial intelligence for web search (AAAI 2000).

[107] Suchacka, G., & Stemplewski, S. (2017). *Application of Neural Network to Predict Purchases in Online Store.* Paper presented at the Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part IV.

[108] Swarnakar, S., Thakur, A., Misra, D., Debopriya, P., Pakira, M., & Roy, S. (2016). Enhanced model of web page prediction using PageRank and markov model. *International Journal of Computer Applications, 140*(7).

[109] Thwe, P. (2014). *Using Markov Model and Popularity and Similarity Based PageRank Algorithm for Web Page Access Prediction.* Paper presented at the International Conference on Advances in Engineering and Technology (ICATE).

[110] Tseng, V. S., Lin, K. W., & Chang, J.-C. (2008). Prediction of user navigation patterns by mining the temporal web usage evolution. *Soft Computing-A Fusion of Foundations, Methodologies and Applications, 12*(2), 157-163.

[111] Verma, A., & Prajapat, B. (2016). User Next Web Page Recommendation using Weight based Prediction. *International Journal of Computer Applications, 142*(11).

[112] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1-37.

[113] Yang, Q., Li, T., & Wang, K. (2004). Building association-rule based sequential classifiers for web-document prediction. *Data mining and knowledge discovery, 8*(3), 253-273.

[114] Yao, Z., Wang, X., & Luan, J. (2017). Using Hidden Markov Model to Predict the Web Users' Linkage. *Journal of Residuals Science & Technology, 14*(3).

[115] Yu, X., Li, M., Paik, I., & Ryu, K. H. (2012). *Prediction of web user behavior by discovering temporal relational rules from web log data.* Paper presented at the International Conference on Database and Expert Systems Applications.

[116] Zack, L., Lamb, R., & Ball, S. (2013). An application of Google's PageRank to NFL rankings. *Involve, a Journal of Mathematics, 5*(4), 463-471.

[117] Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning, 42*(1-2), 31-60.

[118] Zheng, Z., Kohavi, R., & Mason, L. (2001). *Real world performance of association rule algorithms.* Paper presented at the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

[119] Zhu, J., Hong, J., & Hughes, J. G. (2002). Using markov chains for link prediction in adaptive web sites. In *Soft-Ware 2002: Computing in an Imperfect World* (pp. 60-73): Springer.

[120] Zhu, J., Hong, J., & Hughes, J. G. (2002). *Using Markov models for web site link prediction.* Paper presented at the Proceedings of the thirteenth ACM conference on Hypertext and hypermedia.

[121] Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory, 24*(5), 530-536.

PHỤ LỤC 1
MỘT PHẦN MÃ NGUỒN GIẢI PHÁP NÂNG CAO HIỆU QUẢ ĐỘ CHÍNH XÁC CHO DỰ ĐOÁN TRUY CẬP WEB

[Source 1] Duyệt CSDL tuần tự để tạo ra mảng chứa các phần tử khác nhau của CSDL tuần tự

```
1. String s_links="";
2.  for (int i=0; i<arr.length; i++)
3.  {
4.    for (int j=0; j<arr[i].length; j++)
5.        s_links+= arr[i][j]+" ";
6.  }
7.  String[] parts = s_links.split(" ");
8.  Integer[] n1 = new Integer[parts.length];
9.  for(int n =0; n < parts.length; n++)
10. {
11.    n1[n] =  Integer.parseInt(parts[n]);
12. }
13. n1 = removeDuplicates(n1);
14. n1 = Arrays.copyOfRange(n1, 0, n1.length);
```

[Source 2] Tạo ma trận node của CSDL đồ thị từ các phần tử trong CSDL tuần tự

```
1.  String sfile="";
2.  Arrays.sort(n1);
3.  for (int k=0;k<n1.length;k++) {
4.    sfile=sfile+n1[k]+" ";
5.      for (int i=0; i<arr.length; i++)
6.      {
7.          for (int j=0; j<arr[i].length-1; j++)
8.              if (arr[i][j]==n1[k])
9.              {
10.                sfile=sfile+arr[i][j+1]+ " ";
11.             }
12.     }
13.      sfile=sfile.trim();
14.      sfile=sfile +"\n";
15. }
```

[Source 3] Tính toán PageRank cho từng node trong CSDL đồ thị

```
1.  for (int i=0; i<arr_temp.length; i++)
2.  {
3.      for (int j=0; j<arr_temp[i].length; j++) {
4.              for(int n = 0; n < numofLine; n++) {
5.                      if (arr_temp[i][j] == n) {
6.                          System.out.print(arr_temp[i][j].intValue()+ "("+round(globalPageRankValue[n],3)+")" + " " );
7.                         arr_temp[i][j] = round(globalPageRankValue[n],3);
8.                      }
9.              }
10.      }
11.      System.out.println();
12.  }
```

[Source 4] Tính toán trung bình các PageRank cho từng chuỗi dữ liệu tuần tự

```
1. for (int i=0; i<arr_temp.length; i++) {
2.     arr_avg[i] = round(Average_Rows(arr_temp,arr_temp.length, i),3);
3. }
```

[Source 5] Sắp xếp giảm dần theo trung bình PageRank của các chuỗi tuần tự

```
1.  int i, j;
2.  for (i = 0; i < arr.length; i++) {
3.      for (j = 0; j < arr.length; j++) {
4.        if(Double.parseDouble(arr[i].substring(arr[i].length() - 5))>=Double.parseDouble(arr[j].substring(arr[j].length() - 5))) {
5.            String x = arr[i];
6.             arr[i] = arr[j];
7.            arr[j] = x;
8.         }
9.      }
10. }
```

PHỤ LỤC 2

MỘT PHẦN MÃ NGUỒN GIẢI PHÁP NÂNG CAO HIỆU QUẢ THỜI GIAN CHO DỰ ĐOÁN TRUY CẬP WEB

[Source 6] Một phần mã nguồn giải pháp nâng cao hiệu quả thời gian cho dự đoán truy cập Web

```
1. for (int i=0;i<arr_sequence.length;i++)
2.     if(check_contain(arr_sequence[i].split(" "),arr_query)==true)
3.     {
4.           if(arr_sequence[i]!=null) {
5.               str_contain_query = str_contain_query + arr_sequence[i]+"_";
6.               k ++;
7.           }
8.     }
9.      String [] arr_contain_query =str_contain_query.split("_");
10.     String [] seq_selected = new String[k];
11.     String SD_OK = "";
12.  for (int i=0;i<k;i++)
13.      if(arr_contain_query[i].substring(arr_contain_query[i].length() - query.length()).contains(query)==false
14.       || (arr_contain_query[i].substring(arr_contain_query[i].length() - query.length()).contains(query)==true
15.      &&  sub_count(arr_contain_query[i],query)>1))
16.       {
17.            System.out.println(arr_contain_query[i].replace(" ", " -1 ")+" -2");
18.            SD_OK+= arr_contain_query[i].replace(" ", " -1 ")+" -2\n";
19.       }
```

PHỤ LỤC 3

CHI TIẾT GIẢI THUẬT TÍNH TOÁN SONG SONG PAGERANK

```
Procedure Parallel_PageRank
Begin
1.  For i ← 0 to (iterations - 1) do
2.  Begin
3.      For j ← 0 to (n - 1) do
4.          localPR[j] ← 0;     danglingContrib ← 0;
5.      Iterator it = adjMatrix.entrySet().iterator();
6.      While (it.hasNext())
7.          Begin
8.              List pair ← it.next();
9.              If pair.getValue() = null Then    //If it is a dangling node,
10.             danglingContrib ←  danglingContrib + globalPR[pair.getKey()]/n;
11.            Else
12.              Begin
13.            current_size = pair.getValue().size();  iter = pair.getValue().iterator();
14.              While (iter.hasNext())// For each outbound link for a node
15.                  Begin
16.                      node ← iter.next(); temp ← globalPR[node];
17.                      temp  ←  temp + globalPR[pair.getKey()] / current_size;
18.                      localPR[node] = temp;
19.                  End //While (iter.hasNext())
20.                    End //If… Else… Then
21.             End // While (it.hasNext())
22.     tempSend[] ← new double[1];
23.     tempRecv[] ← new double[1];
```

```
24.      tempSend[0] ← danglingContrib;
25.      Call Allreduce(tempRecv, tempSend, MPI.SUM);
26.      Call Allreduce(localPR, globalPR, n,MPI.SUM);
27.      If rank = 0 Then
28.         Begin
29.            For k ← 0 to n do
30.               Begin
31.                  globalPR[k] ← globalPR[k]  + tempRecv[0];
32.                  globalPR[k] ← df * globalPR[k] + (1 - df) * (1/n);
33.               End
34.         End
35.     Call Bcast(globalPR, n);
36.   End // For i ← 0 to  (iterations - 1) do
End
```

Ý KIẾN CỦA NGƯỜI HƯỚNG DẪN 1	NGƯỜI THỰC HIỆN
(Ký ghi rõ họ tên)	(Ký ghi rõ họ tên)
TS. TÂN HẠNH	**NGUYỄN THÔN DÃ**

Ý KIẾN CỦA NGƯỜI HƯỚNG DẪN 2

(Ký ghi rõ họ tên)

TS. PHẠM HOÀNG DUY