BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIÊN THÔNG

TÓM TẮT LUẬN ÁN

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

NCS: NGUYỄN THÔN DÃ

NGƯỜI HƯỚNG DẪN KHOA HỌC: **TS. TÂN HẠNH**
TS. PHẠM HOÀNG DUY

HÀ NỘI, NĂM 2020

Công trình hoàn thành tại:
Học viện Công nghệ Bưu chính Viễn thông

Người hướng dẫn khoa học:
TS. Tân Hạnh

TS. Phạm Hoàng Duy

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng cấp Học viện tại Học viện Công nghệ Bưu chính Viễn thông, 122 Hoàng Quốc Việt, Hà Nội.
Vào lúc:

Có thể tìm hiểu luận án tại:
Thư viện Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

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. Hơn nữa, 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.

Mục tiêu và phạm vi nghiên cứu

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:

(1) 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.

(2) Tìm một mô hình cơ sở dữ liệu phù hợp để hỗ trợ cho dự đoán hành vi truy cập Web.

(3) Tìm giải pháp tốt hơn để nâng cao tính chính xác cho dự đoán hành vi truy cập Web.

(4) Tìm giải pháp tốt hơn để giảm thời gian thực thi dự đoán hành vi truy cập Web.

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).

Ý nghĩa và đóng góp

Khai phá dữ liệu tuần tự cho dự đoán truy cập Web là một trong những nghiên cứu quan trọng trong khai phá dữ liệu. Chẳng hạn như dự đoán hành vi truy cập Web của người học các lớp học trực tuyến, hành vi truy cập bất hợp pháp của tội phạm mạng, hành vi của khác hàng trên các Website thương mại điện tử. Nhiều công trình đã thực hiện và đạt được những kết quả nhất định về độ chính xác và hiệu năng về thời gian dự đoán. Tuy nhiên, để

dự đoán truy cập Web hiệu quả, cần đề xuất các giải pháp tốt hơn về độ chính xác cũng như về thời gian.

Các đóng góp của luận án gồm:

(1) Đề xuất một giải pháp để thiết kế và xây dựng cơ sở dữ liệu tuần tự cho dự đoán truy cập Web.

(2) Đề xuất một giải pháp để làm giảm thời gian dự đoán cho dự đoán truy cập Web.

(3) Đề xuất một giải pháp để tăng độ chính xác cho dự đoán truy cập Web.

(4) Đề xuất một mô hình kết hợp giữa tăng độ chính xác và giảm thời gian dự đoán.

Bố cục 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.

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

1.1. Giới thiệu

Để dự đoán truy cập Web, nhiều nghiên cứu sử dụng các tiếp cận dựa trên máy học. Chẳng hạn, một số các công trình khoa học dùng phương pháp các như Association Rules, Sequential Pattern, Sequential Rules, Markov và các phương pháp lai. Độ chính xác dự đoán được xác định bằng công thức:

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

Trong đó

Accuracy: Độ chính xác của dự đoán

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

|*sequences*|: Số lượng chuỗi dự đoán

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.

Định nghĩa 1.2

Cơ sở dữ liệu tuần tự truy cập Web $SD = \{s_1, s_2, ..., s_N\}$ 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 $S_{query} = \langle page_1, page_2, \dots page_m \rangle$, $S_{query} \in Seq$ 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 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.

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+).

1.3. Các phương pháp phổ biến

Theo F. Khalil và các đồng sự, 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.

*** 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. 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.
  - ✓ 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.
  - ✓ 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.

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.

- ***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:*** 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+ 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+. Chi tiết của mô hình CPT+ được cải tiến từ CPT theo ba chiến lược: Frequent Subsequence Compression (FSC), Simple Branches Compression (SBC), Prediction with improved Noise Reduction (PNR).

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.
- ***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, 47] . Vì thế cần đề xuất các giải pháp để làm tăng tốc độ thời gian dự đoán mà độ chính xác vẫn bảo toàn.
  - ✓ 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.

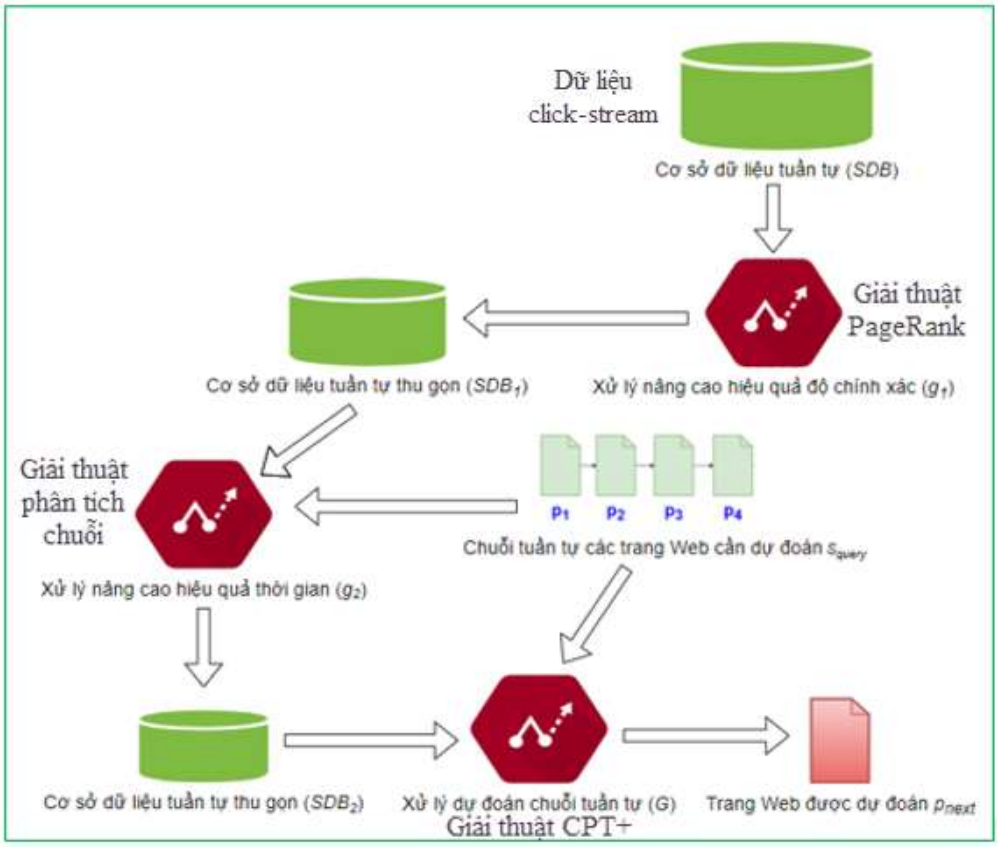
1.4.4. Tổng hợp so sánh các phương pháp dự đoán chuỗi dữ liệu tuần tự

Trên tập dữ liệu BMS, 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.

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

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 1.1.***



Hình 1.1 *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 Page Rank.

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:

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

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.

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

1.6. 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].

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.

2.2. Hạn chế của dự đoán truy cập tuần tự trên Web Log

2.2.1. Hạn chế về không gian dự đoán

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. 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. 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.

2.2.2. Hạn chế về thời gian dự đoán

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. 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

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 [101].

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.

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. 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...

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. [86]

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. [86] 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. Đa số các máy chủ Web hỗ trợ dưới dạng tùy chọn mặc định, định dạng tập tin nhật ký chung (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.

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

2.3.4.1. Ý nghĩa của việc xây dựng cơ sở dữ liệu tuần tự

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.

2.3.4.2. Giải thuật chuẩn hóa cơ sở dữ liệu tuần tự từ cơ sở dữ liệu Web Log

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.

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 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)

2.3.6. 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.

**** 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 2.1 *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

**** Các kết quả thử nghiệm:***

Bảng 2.2 *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

Kết quả thử nghiệm 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 tốt 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].

2.3.7. 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 2.3 *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
Website 2 *palmviewsanibel.com* [2]	4217568	9678
Website 3 *devqa.robotec.co.il* [3]	2527429	312
Website 4 *inees.org* [4]	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; (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.

2.3.7. Kết luận chương 2

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

[1] Truy cập ngày 22/8/2017

[2] Truy cập ngày 22/8/2017

[3] Truy cập ngày 23/8/2017

[4] Truy cập ngày 23/8/2017

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[5]. 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 Page Rank 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. Ý tưởng của giải pháp sử dụng Page Rank để nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web

Một số lý do tính toán Page Rank 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.

(2) 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 và HITS để khai thác cấu trúc Web nhằm nắm bắt tầm quan trọng của một trang Web.

(3) Về bản chất, PageRank diễn giải một siêu liên kết từ trang pageA đến trang pageB dưới dạng phiếu bầu.

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

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. Các nhà phát triển giải thuật PageRank đưa ra một công thức để tính chỉ số Page Rank 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)

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

3.4.1. Phương pháp thực hiện

Giả sử một cơ sở dữ liệu tuần tự *SD* có *N* chuỗi tuần tự.

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

Mỗi cặp liên kết liên tiếp $\{p_i, p_j\}$ theo (trình tự thời gian) trong một chuỗi tuần tự 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 3.2***.

Hình 3.1 *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. 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 *acc1* là độ chính xác của dự đoán chuỗi tuần tự cho cơ sở dữ liệu tuần tự gốc. Tương tự, đặt *acc2* là độ chính xác của dự đoán chuỗi dữ liệu tuần tự của cơ sở dữ liệu được thu gọn. Nếu $acc2 \geq acc1$, 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+.

3.4.2. Giải thuật nâng cao hiệu quả về độ chính xác cho dự đoán truy cập Web

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.

- 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
```

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).

Chi tiết giải thuật tính toán song song PageRank [6] **:**

```
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;
```

[6] Birla, Kai Zhen (Distributed System, CSCI B534, Fall 2016, github.com/cocosci/MPIPagerank)

```
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
```

- 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.

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
```

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

Để đánh giá độ chính xác và thời gian của giải pháp đề xuất, nghiên cứu sinh đã thực hiện các thử nghiệm và đánh giá kết quả thu được.

3.5.2. 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. Dữ liệu

Nghiên cứu sinh đã chọn 3 tập dữ liệu MSNBC, FIFA, KOSARAK 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.

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.

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).

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 Page Rank

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%.

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 bào 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 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.

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. 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. Ý tưởng của giải pháp nâng cao hiệu quả về thời gian cho dự đoán truy cập Web

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

Phương pháp 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

Các tập dữ liệu cho được sử dụng bao gồm BMS, FIFA, SIGN, KOSARAK, BIBLE [7]. 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 [81], PPM [23], AKOM [90].

Về thời gian huấn luyện (training time), kết quả nghiên cứu 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, 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 [48], All-K-Order-Markov(AKOM) [90], Dependancy Graph (DG) [81] , LZ78 [118], PPM [23], Transition Directed Acyclic Graph(TDAG. 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

[7] philippe-fournier-viger.com/spmf/index.php?link=datasets.php

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. Phương pháp thực hiện

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:

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 (*T1)*

Tính độ đo *Acc1* [47].

Kết quả thu được:

- ✓ Kích cỡ cơ sở dữ liệu tuần tự *SD2_size*.
- ✓ Độ đo Accuracy: *Acc1*.
- ✓ Thời gian thực thi: *T1*.

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 *T1* có nhanh hơn Thời gian thực thi *T* đáng kể hay không?

+ Độ chính xác *Acc1* có tương đương hay cao hơn độ chính xác *Acc*?

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 = " "
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
```

4.4.4. Độ đo đánh giá

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

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. Chi tiết được trình bày trong phần 4.4.

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

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.

4.5.1. Dữ liệu

Đối các tập dữ liệu click-stream, các cơ sở dữ liệu tuần tự được sử dụng trong thử nghiệm: *FIFA*[8], *KOSARAK*[10], *BMS* [10].

Bảng 4.1 *Các tập dữ liệu click-stream được thử nghiệm*

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

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

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* [9] *, inees* [10].

Bảng 4.2 *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)

4.5.2. Kết quả thử nghiệm

4.5.2.1. 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 từng chuỗi dự đoán trong phụ lục 1 với độ tin cậy 99% trên tập dữ liệu FIFA. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 30 lần.

4.5.2.2. 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 từng chuỗi dự đoán trong **Phụ lục 1** với độ tin cậy 99% trên tập dữ liệu KOSARAK. Kết quả cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 30 lần.

4.5.2.3. 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 từng chuỗi dự đoán với độ tin cậy 99% trên tập dữ liệu BMS.

Kết quả thử nghiệm cho thấy thời gian giải pháp cải tiến chạy nhanh hơn trên 100 lần.

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*.

Kết quả thử nghiệm cho thấy thời gian giải pháp cải tiến chạy nhanh hơn khoảng 2.7 lần.

---

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

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

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 từng chuỗi dự đoán với độ tin cậy 99% trên tập dữ liệu *inees*. Kết quả thử nghiệm cho thấy thời gian giải pháp cải tiến chạy nhanh hơn gần 2 lần. Như vậy, 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). Các kết quả thực nghiệm cũng chỉ ra rằng 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

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 $\in$ {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.

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 [11]

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.

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 *(1.1)*) 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).

Theo kết quả thử nghiệm, 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).

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.

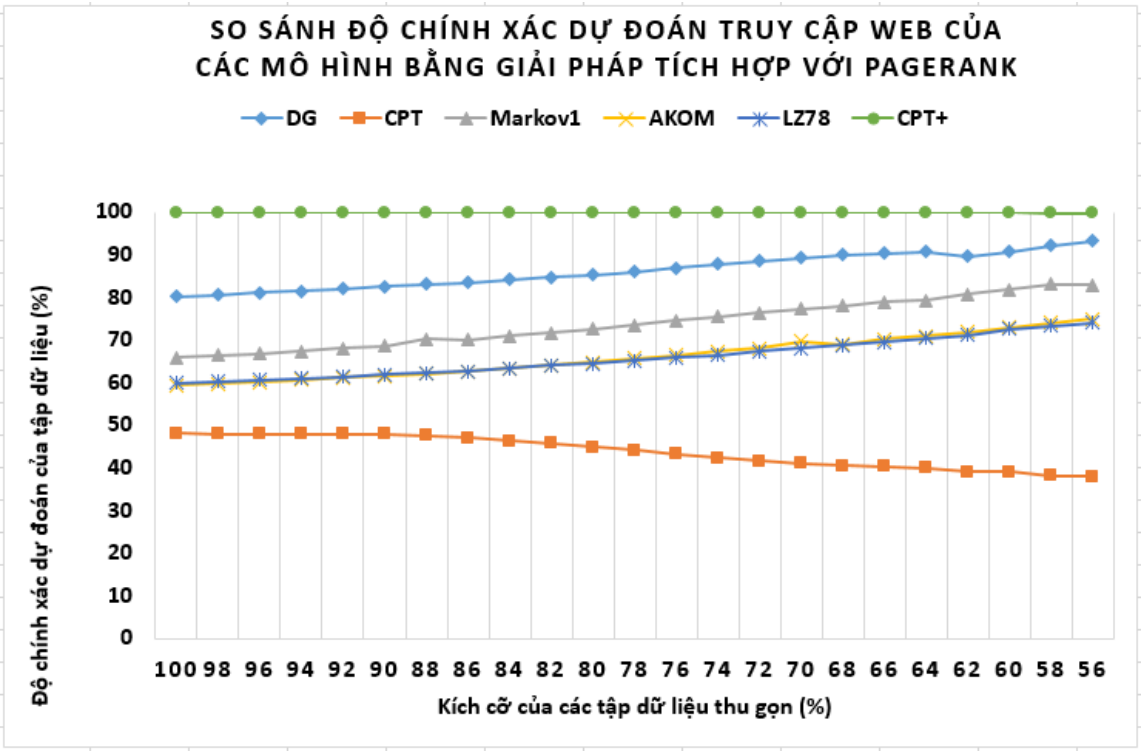
Từ kết quả thực nghiệm trên, ta có cơ sở để khẳng định rằng khi sử dụng tập dữ liệu huấn luyện thu gọn có kích cỡ 66 % (59,400) để tiếp tục cho giai đoạn tiếp là giai đoạn kiểm thử (dự đoán) là rất khả thi.

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

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

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%



Hình 5.1 *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 5.1*** 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. 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].

KẾT LUẬN

1. Đóng góp của luận án

Luận án trình bày 4 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; (4) Giải pháp tích hợp nâng cao độ chính xác và nâng cao hiệu quả về thời gian cho dự đoán truy cập Web.

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 lại:

+ 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 ĐÃ CÔNG BỐ CỦA NGHIÊN CỨU SINH

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

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.

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).

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).

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).

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.

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).

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.

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).

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