TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TP.HCM
KHOA: CÔNG NGHỆ THÔNG TIN

ꕥ-----📖-----ꕥ

BÁO CÁO BÀI TẬP LỚN CUỐI KỲ
Môn học: Khai thác dữ liệu trên Web
Đề tài: Tìm hiểu về Web Usage Mining
Học kỳ 2 (2020 - 2021)

Sinh viên: Vũ Cao Nguyên
MSSV: 18600187
Lớp: 18CK1
Giảng viên: Phạm Trọng Nghĩa

Thành phố Hồ Chí Minh, ngày 28 tháng 08 năm 2021

MỤC LỤC

A. WEB USAGE MINING LÀ GÌ?

Web Usage Mining là một tập hợp con của Khai thác dữ liệu (Data Mining), về cơ bản là trích xuất nhiều loại dữ liệu thú vị khác nhau, sẵn có và có thể truy cập được trên các trang Web, Internet – hay tên gọi chính thức là World Wide Web (WWW). Là một trong những ứng dụng của kỹ thuật khai thác dữ liệu, nó giúp phân tích các hoạt động của người dùng trên các trang web khác nhau và theo dõi chúng trong một khoảng thời gian.

Web Usage Mining là một quá trình gồm ba giai đoạn đó là thu thập và xử lý dữ liệu (Data collection and data preprocessing), phát hiện mẫu (Pattern Discovery) và phân tích mẫu (Pattern Analysis) của dữ liệu web.

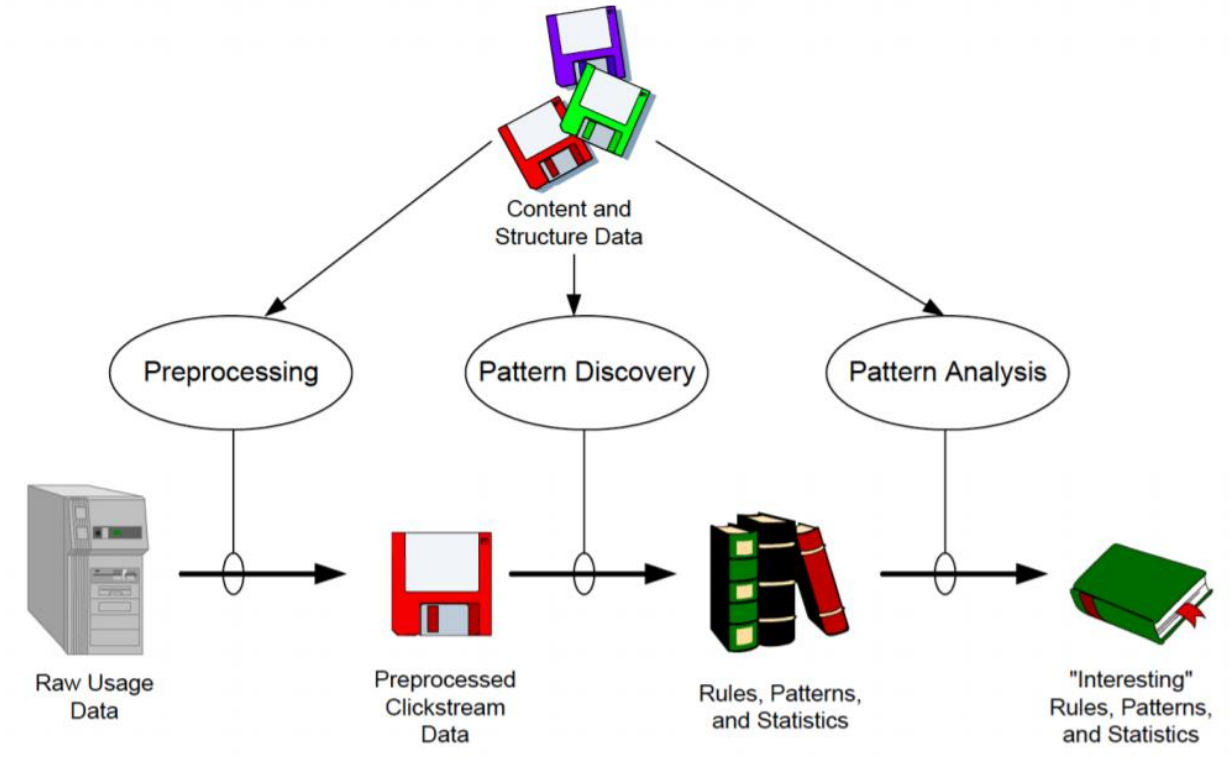

Fig. 1. The Web usage mining (simplified view)

B. QUY TRÌNH CỦA WEB USAGE MINING

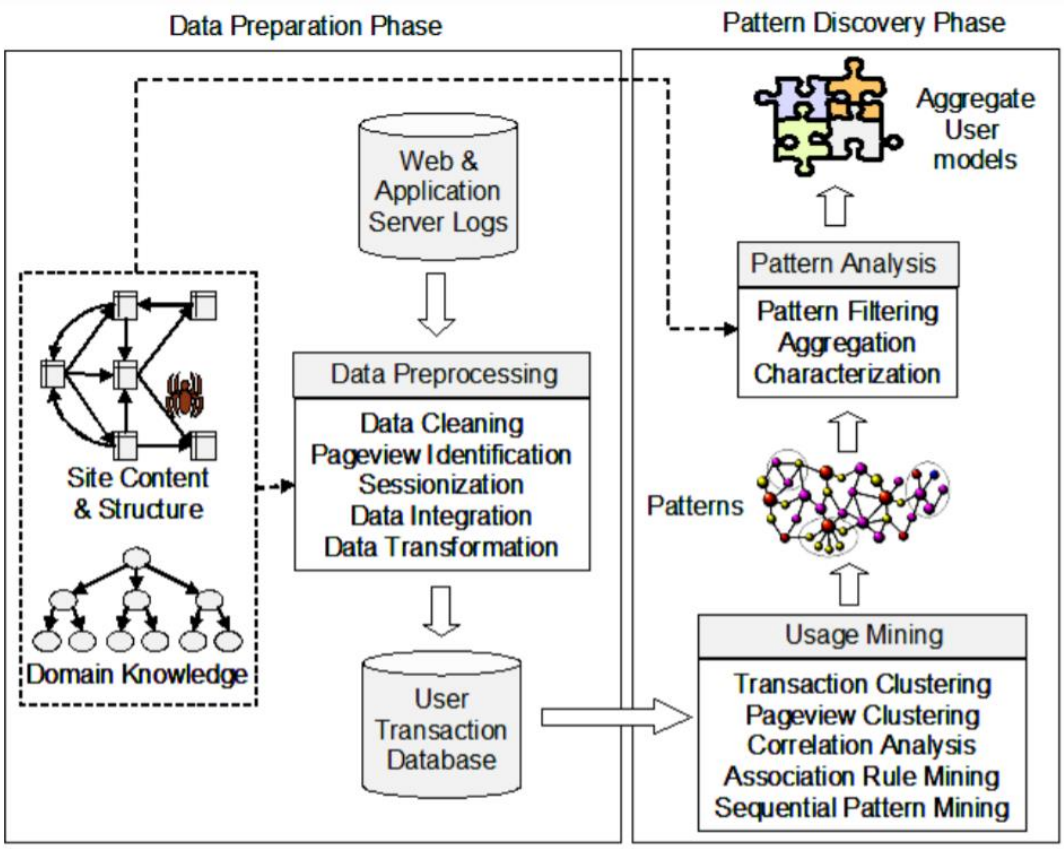

Fig. 2. The Web usage mining process

1. Data collection and data preprocessing

Trong giai đoạn tiền xử lý, dữ liệu được làm sạch và phân vùng thành một tập hợp các giao dịch của người dùng đại diện cho các hoạt động của từng người dùng trong các lần truy cập khác nhau vào trang web. Các nguồn kiến thức khác như nội dung hoặc cấu trúc trang web, cũng như kiến thức miền ngữ nghĩa từ bản thể học trang web (chẳng hạn như danh mục sản phẩm hoặc phân cấp khái niệm), cũng có thể được sử dụng trong quá trình xử lý trước hoặc để nâng cao dữ liệu giao dịch của người dùng. Sự thành công của giai đoạn phân tích mẫu là có liên quan nhiều đến việc chuẩn bị dữ liệu tốt như thế nào nhiệm vụ được thực thi. Điều quan trọng nhất là đảm bảo, mọi sắc thái của nhiệm vụ này đều được xử lý cẩn thận. Việc xử lý giải quyết việc tải dữ liệu, thực hiện kiểm tra độ chính xác, kết hợp dữ liệu với nhau từ các nguồn khác nhau, chuyển đổi dữ liệu thành bắt buộc định dạng và cuối cùng là cấu trúc dữ liệu theo yêu cầu đầu vào của một số thuật toán khai phá dữ liệu. Điều này liên quan đến nhiều giai đoạn như làm sạch

dữ liệu, tính năng trích xuất, giảm tính năng, nhận dạng người dùng, nhận dạng phiên, nhận dạng trang, định dạng và cuối cùng là tóm tắt dữ liệu.

1.1 Sources and Types of Data

Nguồn dữ liệu chính được sử dụng trong Web usage mining là các tệp nhật ký máy chủ (**server log**), bao gồm nhật ký truy cập máy chủ Web (**Web server access logs**) và nhật ký máy chủ ứng dụng (**application server logs**). Các nguồn dữ liệu bổ sung cũng cần thiết cho việc chuẩn bị dữ liệu và khám phá mẫu bao gồm tệp trang web và siêu dữ liệu, cơ sở dữ liệu hoạt động, mẫu ứng dụng và kiến thức miền.

1.1.1 Usage Data

Dữ liệu nhật ký được thu thập tự động bởi máy chủ Web và ứng dụng thể hiện hành vi điều hướng chi tiết của khách truy cập. Nó là nguồn dữ liệu chính trong khai thác sử dụng Web, chẳng hạn như địa chỉ IP của khách truy cập, ngày giờ truy cập, đường dẫn đầy đủ (tệp hoặc thư mục) được truy cập, địa chỉ của người giới thiệu và các thuộc tính khác có thể được đưa vào file log để truy cập Web. Mỗi mục nhập file log (tùy thuộc vào định dạng file log) có thể chứa các trường xác định ngày và giờ của yêu cầu, địa chỉ IP của máy khách, tài nguyên được yêu cầu, các tham số có thể sử dụng để gọi một ứng dụng Web, trạng thái của yêu cầu, phương thức HTTP được sử dụng. Tùy thuộc vào mục tiêu của phân tích, dữ liệu này cần được chuyển đổi và tổng hợp ở các mức độ trừu tượng khác nhau. Mức trừu tượng hóa dữ liệu cơ bản nhất là của một lần xem trang. Số lần xem trang là đại diện tổng hợp của một tập hợp các đối tượng Web góp phần hiển thị trên trình duyệt của người dùng do một hành động người dùng duy nhất (chẳng hạn như nhấp qua). Về mặt khái niệm, mỗi lần xem trang có thể được xem như một tập hợp các đối tượng Web hoặc tài nguyên đại diện cho một "sự kiện người dùng" cụ thể, như: đọc một bài báo, xem trang sản phẩm hoặc thêm sản phẩm vào giỏ hàng.

Ví dụ: mục nhập nhật ký 1 hiển thị người dùng có địa chỉ IP "1.2.3.4" đang truy cập tài nguyên: "/classes/cs589/papers.html" trên máy chủ (maya.cs. depaul.edu). Loại và phiên bản trình duyệt, cũng như thông tin hệ điều hành trên máy khách được ghi lại trong trường tác nhân của mục nhập. Cuối cùng, trường liên kết giới thiệu cho biết rằng người dùng đến vị trí này từ một nguồn bên ngoài: "http://dataminingresources.blogspot.com". Mục nhật ký tiếp theo cho thấy rằng người dùng này đã điều hướng từ "paper.html" (như được phản ánh trong trường

liên kết giới thiệu của mục 2) để truy cập vào một tài nguyên khác: "/classes/cs589/papers/cms-tai.pdf". Mục nhật ký 3 hiển thị một người dùng đã đến tài nguyên "/classes/ds575/papers/hyperlink.pdf" bằng cách thực hiện tìm kiếm trên Google bằng cách sử dụng truy vấn từ khóa: "phân tích siêu liên kết cho khảo sát web". Tất cả các mục từ 4 – 6 đều tương ứng với một lần nhấp qua của người dùng đã truy cập tài nguyên "/classes/cs480/anosystem.html". Mục 5 và 6 là các hình ảnh được nhúng trong tệp "thông báo.html" và do đó, hai yêu cầu HTTP bổ sung được đăng ký dưới dạng lần truy cập trong nhật ký máy chủ tương ứng với các hình ảnh này.

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Fig. 3. Portion of a typical server log

1.1.2 Content Data

Nội dung dữ liệu trong một trang Web là tập hợp các đối tượng và mối quan hệ được chuyển tải đến người dùng. Đối với hầu hết các phần, dữ liệu này bao gồm sự kết hợp của các tài liệu văn bản và hình ảnh. Các nguồn dữ liệu được sử dụng để cung cấp hoặc tạo dữ liệu này bao gồm các trang HTML / XML tĩnh, tệp đa phương tiện, các phân đoạn trang được tạo động từ các tập lệnh và bộ sưu tập các bản ghi từ cơ sở dữ liệu hoạt động. Dữ liệu nội dung trang web cũng bao gồm siêu dữ liệu ngữ nghĩa hoặc cấu trúc được nhúng trong trang web hoặc các trang

riêng lẻ, chẳng hạn như từ khóa mô tả, thuộc tính tài liệu, thẻ ngữ nghĩa hoặc biến HTTP.

1.1.3 Structure Data

Thể hiện quan điểm của nhà thiết kế về tổ chức nội dung trong trang web. Tổ chức này được nắm bắt thông qua cấu trúc liên kết giữa các trang giữa các trang, như được phản ánh thông qua các siêu liên kết. Dữ liệu cấu trúc cũng bao gồm cấu trúc nội trang của nội dung trong một trang.
Ví dụ: cả tài liệu HTML và XML đều có thể được biểu diễn dưới dạng cấu trúc cây trên không gian của các thẻ trong trang.
Cấu trúc siêu liên kết cho một trang web thường được nắm bắt bởi một "sơ đồ trang web" được tạo tự động. Đối với các trang được tạo động, công cụ lập bản đồ trang web phải kết hợp kiến thức nội tại về các ứng dụng và tập lệnh cơ bản tạo ra nội dung HTML hoặc phải có khả năng tạo phân đoạn nội dung bằng cách sử dụng lấy mẫu các tham số được chuyển đến các ứng dụng hoặc tập lệnh đó.

1.1.4 User Data

Cơ sở dữ liệu hoạt động cho trang web có thể bao gồm thông tin hồ sơ người dùng bổ sung. Những dữ liệu đó có thể bao gồm: thông tin nhân khẩu học về người dùng đã đăng ký, xếp hạng của người dùng trên các đối tượng khác nhau như sản phẩm hoặc phim, các giao dịch mua trước đây hoặc lịch sử truy cập của người dùng…
Thông tin còn được thu thập thông qua các biểu mẫu đăng ký hoặc bảng câu hỏi, hoặc có thể được suy ra bằng cách phân tích file log sử dụng Web. Một số dữ liệu này có thể được thu thập ẩn danh miễn là có thể phân biệt được giữa những người dùng khác nhau. Ví dụ: thông tin ẩn danh có trong cookie phía máy khách, nhiều ứng dụng cá nhân hóa yêu cầu lưu trữ thông tin hồ sơ người dùng trước.

1.2 Key Elements of Web Usage Data Pre-Processing

Các tác vụ cấp cao được yêu cầu trong xử lý trước, dữ liệu sử dụng bao gồm kết hợp và đồng bộ hóa dữ liệu từ nhiều tệp nhật ký, làm sạch dữ liệu, nhận dạng số lần xem trang, nhận dạng người dùng, nhận dạng phiên (hoặc phân loại), nhận dạng tập và tích hợp dữ liệu luồng nhấp chuột với các nguồn dữ liệu khác như thông tin nội dung hoặc ngữ nghĩa, cũng như thông tin người dùng và sản phẩm từ cơ sở dữ liệu hoạt động. Thông tin có sẵn trên web là không đồng nhất và không có cấu trúc. Do đó, giai đoạn tiền xử lý là điều kiện tiên quyết để phát hiện ra các mẫu. Tiền xử lý dữ liệu đưa ra một số thách thức riêng dẫn đến nhiều thuật

toán và kỹ thuật heuristic cho các tác vụ tiền xử lý như hợp nhất và dọn dẹp, nhận dạng người dùng và phiên, v.v. Nhiều công trình nghiên cứu khác nhau được thực hiện trong lĩnh vực tiền xử lý này để nhóm các phiên và giao dịch, đó là được sử dụng để khám phá các mẫu hành vi của người dùng.

1.2.1 Data Fusion and Cleaning

Làm sạch dữ liệu là một quá trình xóa các mục không liên quan như tệp jpeg, gif hoặc tệp âm thanh và các tham chiếu do điều hướng nhện. Chất lượng dữ liệu được cải thiện giúp cải thiện phân tích trên đó. Giao thức Http yêu cầu một kết nối riêng biệt cho mọi yêu cầu từ máy chủ web. Nếu người dùng yêu cầu xem một trang cụ thể cùng với đồ họa và tập lệnh của mục nhập nhật ký máy chủ được tải xuống cùng với file HTML. Làm sạch dữ liệu thường dành riêng cho từng trang web và bao gồm các tác vụ như xóa các tham chiếu không liên quan đến các đối tượng được nhúng có thể không quan trọng cho mục đích phân tích, bao gồm các tham chiếu đến tệp kiểu, tệp đồ họa hoặc tệp âm thanh. Quá trình làm sạch cũng có thể liên quan đến việc loại bỏ ít nhất một số trường dữ liệu (ví dụ: số byte được truyền hoặc phiên bản của giao thức HTTP được sử dụng, v.v.) có thể không cung cấp thông tin hữu ích trong các nhiệm vụ phân tích hoặc khai thác dữ liệu. Việc làm sạch dữ liệu cũng đòi hỏi phải xóa các tham chiếu do điều hướng của trình thu thập thông tin vì một tệp nhật ký điển hình chứa một tỷ lệ phần trăm tham chiếu đáng kể (đôi khi cao tới 50%) do công cụ tìm kiếm hoặc các trình thu thập thông tin khác (hoặc trình thu thập thông tin).
Có ba loại dữ liệu không liên quan hoặc dư thừa cần được làm sạch: tài nguyên phụ trợ được nhúng trong tệp HTML, yêu cầu của rô bốt và yêu cầu lỗi.

- **Accessorial Resources**: Bởi vì giao thức HTTP không có kết nối, yêu cầu của người dùng để xem một trang cụ thể thường dẫn đến một số mục nhật ký vì đồ họa và tập lệnh được tải xuống cùng với tệp HTML. Việc loại bỏ các mục được coi là không liên quan có thể được thực hiện một cách hợp lý bằng cách kiểm tra hậu tố của tên URL. Ví dụ: tất cả các mục nhật ký có hậu tố tên tệp như gif, jpeg, GIF, JPEG, jpg, JPG, css và map đều có thể bị xóa.
- **Robots' requests**: Robot web (còn được gọi là spiders) là công cụ phần mềm quét một trang Web để trích xuất nội dung của nó. Spider tự động theo dõi tất cả các siêu liên kết từ một trang Web. Các công cụ tìm kiếm như Google định kỳ sử dụng các trình thu thập dữ liệu để lấy tất cả các trang từ một trang Web để cập nhật các chỉ mục tìm kiếm của chúng. Để

xóa yêu cầu của robots, ta có thể tìm kiếm tất cả các máy chủ đã yêu cầu trang "robots.txt".

- **Error's requests**: Error's requests là vô ích cho quá trình khai thác. Chúng có thể được loại bỏ bằng cách kiểm tra trạng thái của yêu cầu. Ví dụ: nếu trạng thái là 404, nó cho thấy rằng tài nguyên được yêu cầu không tồn tại. Sau đó, mục nhập nhật ký này trong nhật ký có thể được gỡ bỏ.

1.2.2 Pageview Identification

Việc xác định số lần xem trang phụ thuộc nhiều vào cấu trúc nội bộ của trang web, cũng như vào nội dung trang và kiến thức về miền cơ bản của trang web. Xác định tập hợp các tệp / đối tượng / tài nguyên Web đại diện cho một "sự kiện người dùng" cụ thể tương ứng với một nhấp qua (ví dụ: xem trang sản phẩm, thêm sản phẩm vào giỏ hàng). Trong một số trường hợp, có thể tốt nếu xem xét số lần xem trang ở cấp độ tổng hợp cao hơn, ví dụ: chúng có thể tương ứng với nhiều sự kiện của người dùng liên quan đến cùng một danh mục khái niệm, chẳng hạn như việc mua một sản phẩm trên trang web thương mại điện tử trực tuyến.

1.2.3 User Identification

Xác định người dùng cá nhân truy cập vào một trang web là một bước quan trọng trong khai thác sử dụng web. Các phương pháp khác nhau sẽ được tuân theo để xác định người dùng. Phương pháp đơn giản nhất là gán id người dùng khác nhau cho địa chỉ IP khác nhau. Nhưng trong máy chủ Proxy, nhiều người dùng đang chia sẻ cùng một địa chỉ và cùng một người dùng sử dụng nhiều trình duyệt. Sự cố bộ nhớ đệm có thể được khắc phục bằng cách ấn định thời gian hết hạn ngắn cho các trang HTML để thực thi trình duyệt truy xuất mọi trang từ máy chủ. Trong trường hợp không có cơ chế xác thực, cách tiếp cận phổ biến nhất để phân biệt giữa các khách truy cập là sử dụng cookie phía máy khách. Tuy nhiên, không phải tất cả các trang web đều sử dụng cookie và do lo ngại về quyền riêng tư (bị người dùng vô hiệu hóa). Chỉ riêng địa chỉ IP để xác định người dùng là không đủ, trên thực tế, nếu có hai lần xuất hiện của cùng một địa chỉ IP (cách nhau một khoảng thời gian đủ) có thể tương ứng với hai người dùng khác nhau. Vì vậy có thể xác định chính xác người dùng duy nhất thông qua sự kết hợp của địa chỉ IP và các thông tin khác như tác nhân người dùng và liên kết giới thiệu.

Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser.	Can be turned off by users.
Software Agents	Program loaded into browser and sends back usage data.	High	Accurate usage data for a single site.	Likely to be rejected by users.

Ví dụ: Ở bên trái, hình mô tả một phần của tệp nhật ký được xử lý trước một phần (các dấu thời gian chỉ được đưa ra dưới dạng giờ và phút). Sử dụng kết hợp các trường IP và Agent trong tệp nhật ký, chúng tôi có thể phân vùng nhật ký thành các bản ghi hoạt động cho ba người dùng riêng biệt (được mô tả ở bên phải).

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Fig. 4. Example of user identification using IP + Agent

1.2.4 Sessionization

Sessionization là quá trình phân đoạn bản ghi hoạt động người dùng của từng người dùng thành các phiên, mỗi phiên đại diện cho một lượt truy cập vào trang web. Các trang web không được hưởng lợi từ thông tin xác thực bổ sung từ người dùng và không có các cơ chế như id phiên được nhúng phải dựa vào các phương pháp heuristics để phân loại. Mục tiêu của phương pháp **sessionization heuristics** là tái tạo lại từ dữ liệu luồng nhấp chuột, chuỗi hành động thực tế được thực hiện bởi một người dùng trong một lần truy cập vào trang web. Có hai loại phương pháp **sessionization heuristics** cơ bản: **Time-oriented** hoặc **Structure-oriented**

- **Time-oriented**: áp dụng ước tính thời gian chờ toàn cục hoặc cục bộ để phân biệt giữa các phiên liên tiếp.
- **Structure-oriented**: sử dụng cấu trúc trang web tĩnh hoặc cấu trúc liên kết ngầm được ghi lại trong các trường liên kết giới thiệu của nhật ký máy chủ

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 5. Example of sessionization with a time-oriented heuristic

Episode identification: có thể được thực hiện như một bước cuối cùng trong quá trình xử lý trước dữ liệu dòng nhấp để tập trung vào các tập hợp con có liên quan của số lần xem trang trong mỗi phiên người dùng. Một Episode là một tập hợp con hoặc chuỗi con của session bao gồm các lần xem trang liên quan đến ngữ nghĩa hoặc chức năng.

1.2.5 Path Completion

Một nhiệm vụ tiền xử lý quan trọng khác thường là được thực hiện sau khi quá trình sessionization là quá trình hoàn thành đường dẫn (**path completion**). Phía máy khách hoặc phía proxy bộ nhớ đệm thường có thể dẫn đến việc thiếu các tham chiếu truy cập đến các trang đó hoặc các đối tượng đã được lưu vào bộ nhớ

đệm. Ví dụ: Nếu người dùng quay lại trang A trong cùng một session, lần truy cập thứ hai vào A có thể sẽ dẫn đến việc xem trước đó đã tải xuống phiên bản A đã được lưu vào bộ nhớ đệm ở phía máy khách, và do đó, không có yêu cầu nào được thực hiện đối với máy chủ. Điều này dẫn đến tham chiếu thứ hai đến A không được ghi vào nhật ký máy chủ.

Missing references: do lưu vào bộ nhớ đệm có thể được suy ra thông qua việc hoàn thành đường dẫn dựa trên kiến thức về cấu trúc trang web và thông tin liên kết giới thiệu từ nhật ký máy chủ. Ví dụ về Missing references như hình 6: Sau khi đến trang E, người dùng đã lùi (ví dụ: sử dụng nút "quay lại" của trình duyệt) đến trang D và sau đó là B mà từ đó họ đã điều hướng đến trang C. Các tham chiếu quay lại đến D và B không xuất hiện trong tệp nhật ký vì các trang này được lưu trong bộ nhớ cache ở phía máy khách.

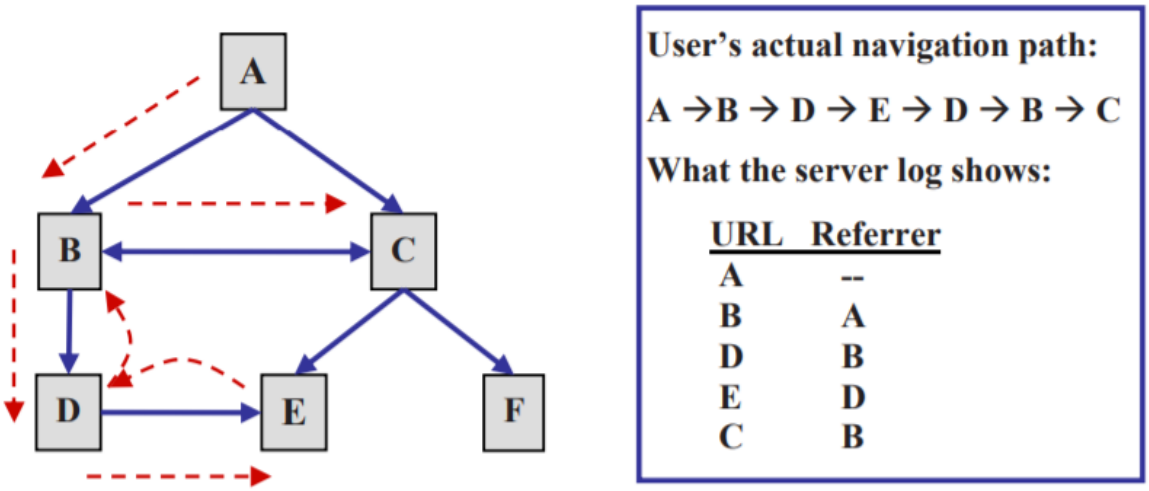

Fig. 6. Missing references due to caching

1.2.6 Data Integration

Cung cấp khuôn mẫu hiệu quả nhất cho việc khám phá mẫu, dữ liệu từ nhiều nguồn khác phải được tích hợp với dữ liệu. Đồng bộ hóa dữ liệu từ nhiều nhật ký máy chủ. Tích hợp thương mại điện tử và dữ liệu máy chủ ứng dụng, siêu dữ liệu, dữ liệu nhân khẩu. Một số ví dụ về các chỉ số đó bao gồm tần suất hoặc giá trị tiền tệ của các lần mua hàng, quy mô trung bình của các giỏ hàng trên thị trường, số lượng các mặt hàng khác nhau được mua, số lượng các danh mục mặt hàng khác nhau được mua, lượng thời gian dành cho các trang hoặc phần của trang web, ngày trong tuần và thời gian trong ngày khi một hoạt động nhất định xảy ra, phản hồi cho các đề xuất và thông tin đặc biệt trực tuyến, v.v.

2. Pattern Discovery

Khám phá mẫu tập trung vào việc khám phá các mẫu từ các phần trừu tượng được tạo ra do kết quả của giai đoạn tiền xử lý. Nó tập trung vào việc áp dụng các phương pháp và kỹ thuật khác nhau được phát triển từ một số lĩnh vực như khai thác dữ liệu, học máy, thống kê và nhận dạng mẫu.

2.1 Statistical Analysis

Phân tích thống kê tập trung vào việc phân tích dữ liệu được tổng hợp theo các đơn vị xác định trước như ngày, phiên, khách truy cập để có được kiến thức về hành vi của người dùng. Chủ yếu ba loại phân tích thống kê (frequency, mean, median) là được thực hiện trên các phiên và kết quả phân tích hiển thị các trang được truy cập thường xuyên nhất, thời gian xem trung bình hoặc độ dài của trang. Những kết quả này trở nên hữu ích trong việc cải thiện hiệu suất hệ thống hoặc tăng cường bảo mật.

2.2 Association Rules

Các quy tắc kết hợp được sử dụng để tìm mối tương quan giữa các trang web thường xuyên xuất hiện cùng nhau trong trình duyệt web của người dùng. Thuật toán Apriori là thuật toán phổ biến nhất thể hiện sự đồng xuất hiện thường xuyên của các trang web với nhau. Các quy tắc này giúp cung cấp các đề xuất cho người dùng khi họ truy cập các trang web và nhà thiết kế web để tái cấu trúc trang web.

2.3 Clustering

Clustering là một kỹ thuật khai thác dữ liệu để nhóm một tập hợp các mục có các đặc điểm tương tự lại với nhau. Có hai loại **Clusters** đó là: **user clusters** và **page clusters**. Phân cụm các bản ghi người dùng (phiên hoặc giao dịch) là một trong những nhiệm vụ phân tích được sử dụng phổ biến nhất trong khai thác sử dụng Web và phân tích Web. Phân nhóm người dùng có xu hướng thiết lập các nhóm người dùng có các kiểu duyệt web tương tự. Những kiến thức như vậy đặc biệt hữu ích cho việc phỏng đoán nhân khẩu học của người dùng để thực hiện phân đoạn thị trường trong các ứng dụng thương mại điện tử hoặc cung cấp nội dung Web được cá nhân hóa cho những người dùng có cùng sở thích. Phân tích sâu hơn về các nhóm người dùng dựa trên các thuộc tính nhân khẩu học của họ (ví dụ: tuổi, giới tính, mức thu nhập, v.v.) có thể dẫn đến việc khám phá thông tin kinh doanh có giá trị.

2.4 Classification

Phân loại (**classification**) là một quá trình học tập có giám sát, bởi vì việc học được thúc đẩy bởi việc phân công các cá thể cho các lớp trong khóa đào tạo dữ liệu. Các thuật toán học quy nạp được sử dụng cho việc thực hiện phân loại gồm: **decision trees**, **naive Bayesian classifiers**, **k-nearest neighbor classifiers** và **support vector machine**. **Classification** chủ yếu thực hiện việc phân loại tài liệu tự động. Trong khai thác sử dụng web, ứng dụng thuật toán phân loại trên nhật ký máy chủ có thể dẫn đến việc phát hiện các mẫu thú vị chẳng hạn như 40% người dùng truy cập trang web tin tức thuộc nhóm tuổi 30-35 năm.

2.5 Sequential Patterns

Trong khai thác sử dụng web, sequential patterns được sử dụng để khám phá các session được tìm thấy trong một trình tự. Chúng bao gồm chuỗi các mục thường xuyên xảy ra theo một thứ tự cụ thể. Thuật toán MIDAS (Khai thác dữ liệu Internet cho chuỗi liên kết) được sử dụng phổ biến nhất để tìm kiếm các mẫu tuần tự cung cấp hành vi tiếp thị thông minh cho kịch bản thương mại điện tử.

3. Pattern Analysis

Phân tích mẫu (**pattern analysis**) là giai đoạn cuối cùng trong khai thác sử dụng web. Các patterns được khai thác không thích hợp để diễn giải và phán đoán. Vì vậy, điều quan trọng là phải lọc ra các quy tắc hoặc patterns không tốt ra khỏi tập hợp được tìm thấy trong giai đoạn pattern discovery. Trong giai đoạn này, các công cụ được cung cấp để tạo điều kiện thuận lợi cho việc chuyển đổi thông tin thành kiến thức. Cơ chế truy vấn tri thức như SQL là phương pháp Pattern Analysis phổ biến nhất. Một phương pháp khác là tải dữ liệu sử dụng vào một khối dữ liệu để thực hiện các hoạt động OLAP (một công nghệ cơ sở dữ liệu đã được tối ưu hóa cho truy vấn và báo cáo, thay vì xử lý các giao dịch). Kết quả của giai đoạn phân tích mẫu được sử dụng trong các ứng dụng khác nhau như cải thiện hiệu suất hệ thống, trang web sửa đổi, cá nhân hóa, thương mại điện tử, v.v. Các mẫu có thể được phân tích bằng cách sử dụng các kỹ thuật sau được mô tả dưới đây:

3.1 Knowledge Query Mechanism

Ngôn ngữ truy vấn có cấu trúc (SQL) là ngôn ngữ được sử dụng phổ biến nhất cho cơ chế truy vấn kiến thức. Ngôn ngữ này được áp dụng để trích xuất các mẫu hữu ích từ các mẫu đã khám phá.

3.2 OLAP/Visualization tools

Được phân tích bằng cách sử dụng các công cụ OLAP trong đó các dữ kiện đã khám phá được đưa vào các khối dữ liệu để thực hiện các hoạt động OLAP khác nhau như cuộn lên và đi sâu và các dữ kiện thú vị được truy xuất. OLAP cung cấp một khung tích hợp để phân tích cho phép thay đổi các mức tổng hợp. Đầu ra của các truy vấn OLAP hoạt động như một đầu vào cho các công cụ khai thác dữ liệu hoặc trực quan hóa dữ liệu. Các mẫu đồ thị hoặc gán màu cho các giá trị khác nhau được sử dụng như một kỹ thuật trực quan hóa cho cùng một mục đích.

3.3 Intelligent Agents

Nhiều tác nhân khác nhau cũng được tạo ra để giúp kiểm tra các mẫu trong khai thác sử dụng web. Các tác nhân này thực hiện công việc phân tích các mẫu được phát hiện.

C. KẾT LUẬN

Khai thác sử dụng web (**Web Usage Mining**) được sử dụng trong nhiều lĩnh vực khác nhau như kinh doanh điện tử, e-CRM, giáo dục điện tử, tin sinh học và thư viện kỹ thuật số, v.v. Thông tin duy nhất mà người dùng để lại khi họ truy cập bất kỳ trang web nào là về kiểu truy cập của họ. Khai thác sử dụng web sử dụng thông tin này để khai thác thông tin mong muốn và cung cấp thông tin đó cho người dùng một cách hiệu quả và hiệu quả. Xử lý trước nội dung và cấu trúc cho phép dữ liệu thô cũng được xử lý trước theo các kích thước này. Các mẫu được phát hiện bằng cách sử dụng các kỹ thuật khác nhau như phân tích thống kê, phân cụm, quy tắc kết hợp, v.v. Sự tham gia của các tác nhân thông minh và cơ chế truy vấn kiến thức giúp cải thiện hiệu quả của việc phân tích mẫu.

D. TÀI LIỆU THAM KHẢO

[1] Web Data Mining, 2nd Edition: Exploring Hyperlinks, Contents, and Usage Data.

[2] A Review on Pattern Discovery Techniques of Web Usage Mining.

[3] Web Usage Mining Tools & Techniques: A Survey.

[4] An Overview on Web Usage Mining.

[5] Research and Development of Data Preprocessing in Web Usage Mining.

[6] Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications.

[7] Data Preprocessing Algorithm for Web Structure Mining.

[8] Web Usage Mining: A Review On Process, Methods and Techniques.

[9] Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.

[10] Analysis of Data Extraction and Data Cleaning in Web Usage Mining.