KỸ THUẬT KHAI THÁC WEB VÀ ỨNG DỤNG
A Survey on Web Mining Techniques and Applications

Văn Thị Thiên Trang
Trường Đại học Kinh tế - Tài chính Thành phố Hồ Chí Minh – UEF, trangvtt@uef.edu.vn

Tóm tắt*: Với lượng thông tin khổng lồ có sẵn trên mạng, World Wide Web là một lĩnh vực màu mỡ cho nghiên cứu khai thác dữ liệu. Khai thác web là một nhánh của khai thác dữ liệu, đó là quá trình khám phá thông tin và tri thức từ khối lượng dữ liệu web đồ sộ. Khai thác web phân thành ba loại chính gồm khai thác cấu trúc web, khai thác nội dung web, khai thác sử dụng web. Tùy thuộc mục đích khai thác và ứng dụng, người dùng có thể sử dụng loại khai thác web tương ứng. Bài báo này trình bày tổng quan về lĩnh vực khai thác web cùng với khảo sát chi tiết về phương pháp và kỹ thuật của từng loại khai thác web, là nền tảng tiền đề cho các nghiên cứu về sau cho từng lĩnh vực ứng dụng cụ thể.*

Từ khóa*: Khai thác dữ liệu, khai thác cấu trúc web, khai thác cấu trúc web, khai thác nội dung web, khai thác sử dụng web.*

1. Giới thiệu

World Wide Web (WWW) là hệ thống tích hợp các trang web mà người dùng có thể truy cập thông qua mạng Internet. Nó là nơi chứa thông tin, tài liệu và tài nguyên của mọi website trên toàn cầu. Ngày nay, WWW ngày càng phổ biến và là phương tiện để truy cập nhiều loại thông tin khác nhau từ các trang web trên khắp thế giới. Vì thông tin trên web gia tăng hàng ngày với tốc độ khổng lồ, thật khó để trích xuất những thông tin hữu ích theo nhu cầu người dùng. Do đó, khai thác web đóng vai trò quan trọng trong việc giải quyết những thách thức này. Nhìn chung, khi người dùng sử dụng web, các vấn đề đáng quan tâm là tìm kiếm thông tin có liên quan, thu thập kiến thức từ dữ liệu có sẵn trên web, tạo ra kiến thức mới bằng cách sử dụng thông tin có sẵn trên web.

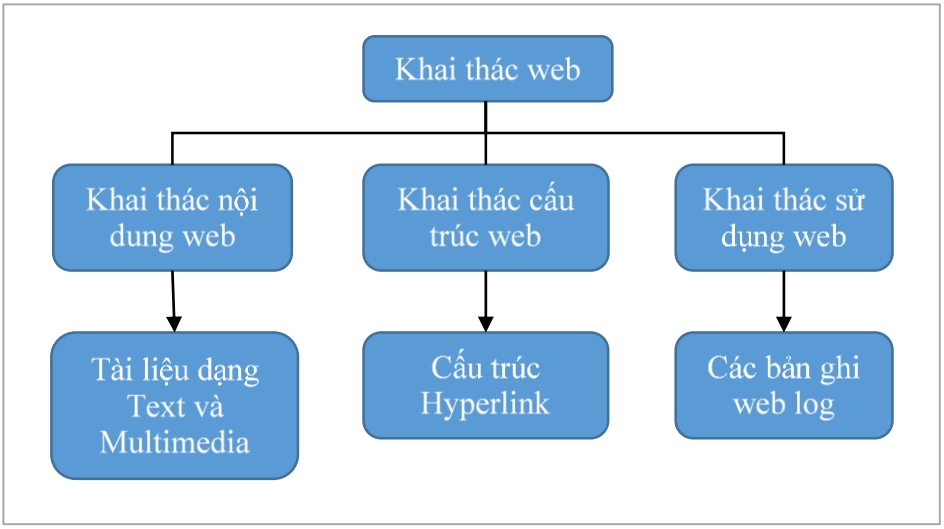
Khai thác web là một phân ngành của khai thác dữ liệu, liên quan đến thông tin truy xuất từ internet. Khai thác web là quá trình sử dụng các công nghệ và kỹ thuật khai thác dữ liệu để rút trích tri thức từ dữ liệu web. Dữ liệu web bao gồm ba loại chính, đó là các tài liệu web (documents), các siêu liên kết giữa các tài liệu (hyperlinks), nhật ký sử dụng các trang web của người dùng (web logs).

Cũng như khai thác dữ liệu truyền thống, khai thác web nhằm khám phá và tìm kiếm những mẫu thông tin hữu ích và thú vị từ các tập dữ liệu lớn. Khai thác web thường bao gồm các bước: thu thập dữ liệu, tiền xử lý dữ liệu, khám phá và phân tích tri thức**.**

Về cơ bản, khai thác web được chia làm ba loại gồm khai thác nội dung trang web (Web Content Mining), khai thác cấu trúc web (Web Structure Mining), và khai thác sử dụng web (Web Usage Mining) [1].

(1) Khai thác nội dung trang web: là quá trình rút trích những thông tin hữu ích từ nội dung của các tài liệu Web. Nội dung của một tài liệu web tương ứng với các khái niệm mà tài liệu tìm cách truyền tải tới người dùng. Nội dung này có thể bao gồm văn bản, hình ảnh, video, âm thanh hoặc các bản ghi có cấu trúc như danh sách và bảng biểu. Trong đó, khai thác dữ liệu dạng văn bản thu hút nghiên cứu nhiều hơn các dạng khác.

(2) Khai thác cấu trúc web: là quá trình khám phá và rút trích những thông tin cấu trúc từ các tài liệu Web, trong đó quá trình khai thác được thực hiện ở hai cấp độ tùy thuộc vào loại dữ liệu cấu trúc được sử dụng cho quá trình phân tích, là dữ liệu ở mức siêu liên kết hay mức trang web.



***Hình 1**. Phân loại khai thác web và đối tượng dữ liệu tương ứng.*

(3) Khai thác sử dụng web: là ứng dụng của các kỹ thuật khai thác dữ liệu để khám phá các mẫu sử dụng Web để biết xu hướng truy cập trang web của từng đối tượng người dùng tại các thời điểm khác nhau, từ đó hiểu rõ hơn và đáp ứng nhu cầu của người dùng.

Hình 1 minh họa các loại khai thác web và đối tượng dữ liệu khai thác tương ứng. Qua đó, ta thấy rằng các phạm trù khai thác web là khác nhau dựa trên loại dữ liệu mà chúng khai thác và loại tri thức rút trích được từ đó.

Khai thác web được ứng dụng trong các lĩnh vực như thương mại điện tử, quản trị mối quan hệ khách hàng điện tử, quản lý tri thức, các dịch vụ web, và hiệu suất hệ thống.

Ở các phần tiếp theo, bài báo sẽ trình bày chi tiết từng loại khai thác web cùng với các công nghệ khai thác dữ liệu được áp dụng.

2. Khai thác nội dung web

Khai thác nội dung web là khai thác, trích xuất và tích hợp dữ liệu, thông tin và những kiến thức hữu ích từ nội dung trang web. Khai thác nội dung web tập trung vào nội dung trang web như các văn bản, hình ảnh và các nội dung đa phương tiện đính kèm như audio, video. Khai thác nội dung web có liên quan với khai thác dữ liệu, khai thác văn bản tuy nhiên cần phân biệt vì dữ liệu web chủ yếu là dữ liệu bán cấu trúc và/hoặc phi cấu trúc, trong khi đó khai thác dữ liệu chủ yếu xử lý dữ liệu có cấu trúc. Tùy thuộc vào loại dữ liệu, sẽ áp dụng các kỹ thuật khai thác khác nhau.

2.1. Các kỹ thuật khai thác với dữ liệu phi cấu trúc

Hầu hết dữ liệu sẵn có trên Web đều ở dạng phi cấu trúc, các kỹ thuật khai thác Web phi cấu trúc bao gồm:

Rút trích thông tin: kỹ thuật này rất được quan tâm khi có một lượng lớn văn bản được tích lũy. Có thể áp dụng kỹ thuật này bằng cách chuyển đổi dữ liệu phi cấu trúc sang dạng có cấu trúc, sau đó khai thác thông tin bằng cách sử dụng một số luật [2].

Dò tìm chủ đề: là quá trình tìm các tài liệu web có liên quan đến truy vấn của người dùng bằng cách xác định các trang web có liên quan đến một số chủ đề nhất định, sau đó sử dụng các siêu liên kết để xác định nhóm các trang web có liên quan đến một chủ đề cụ thể [3].

Tóm tắt : là kỹ thuật dùng để làm giảm độ dài của tài liệu web và đưa ra quyết định cho người dùng liệu có nên đọc những tài liệu (trang) web này hay không thay vì phải đọc đoạn đầu tiên để biết tài liệu đó có liên quan đến vấn đề mà người dùng đang quan tâm hay không [4].

Phân loại/phân lớp: là quá trình gán nhãn lớp cho các nội dung web từ bộ nhãn lớp đã xác định trước trong tập dữ liệu, phân lớp nội dung web có thể chia thành hai loại [5]:

Phân lớp Web Page: là quá trình gán nhãn lớp hoặc phân loại cho trang web từ tập các lớp/loại có sẵn. Phân loại trang web khác với quy trình phân loại tài liệu văn bản truyền thống ở các khía cạnh sau: thứ nhất, phân loại tài liệu văn bản truyền thống thường được thực hiện theo kiểu có cấu trúc và nhất quán, trong khi trang web không có thuộc tính này. Thứ hai, các trang web chủ yếu tồn tại ở dạng phi cấu trúc HTML và có cả những thuộc tính khác không tồn tại trong dạng tài liệu văn bản như từ khóa, tiêu đề, thẻ.

Phân lớp Website: có nhiều phương pháp phân lớp website chẳng hạn có thể phân lớp dựa trên nội dung của trang chủ, phân lớp theo các thẻ HTML, hoặc phân lớp dựa trên các thuộc tính cấu trúc liên kết. Phân lớp web page có thể hỗ trợ cho phân lớp website vì nếu biết chủ đề của web page, chúng ta cũng suy ra được chủ đề của website vì website là một tập hợp rộng chứa nhiều web page nhỏ khác nhau. Phạm vi hoạt động của website lớn và bao hàm cả webpage bên trong.

Phân cụm (Clustering): là một kỹ thuật được sử dụng để nhóm các đối tượng tương tự nhau bao gồm một số nội dung Web, được phân cụm dựa trên các đặc điểm hoặc tham số nhất định. Phân cụm web có thể là một trong các loại sau [3][6]:

Phân cụm Web Page (trang tài liệu): các trang web được nhóm dựa trên những nội dung có liên quan nhau, thông tin này rất hữu ích trong các công cụ tìm kiếm và phương pháp truy lục thông tin, giúp tăng khả năng tiếp cận và phát triển nội dung.

Phân cụm đối tượng web: gom nhóm nội dung có liên quan để phục vụ truy vấn của người dùng, các đối tượng phân cụm có thể bao gồm tệp văn bản, hình ảnh, video và bản âm thanh.

Phân cụm Web Site: là quá trình nhằm nhóm các website có các đặc điểm tương tự nhau. Tác vụ này gặp phải các thách thức như việc rút trích nội dung văn bản từ một số trang web là nhiệm vụ phức tạp và đòi hỏi nhiều bước tiền xử lý; hơn nữa, các nội dung đa phương tiện như hình ảnh, âm thanh và video cần các kỹ thuật mới và phải thực hiện không tốn quá nhiều thời gian và chi phí xử lý.

Trực quan hóa thông tin: Các công cụ trực quan có thể giúp chúng ta hiểu rõ hơn về các nội dung web. Do đó, cần phát triển công cụ có khả năng biểu diễn các đối tượng web dưới dạng đồ họa như thời gian sử dụng web, luồng nhấp chuột của người dùng vào trang web, mối quan hệ giữa các website, tương quan giữa các trang web… Về mặt khái niệm, có nhiều công cụ trực quan hỗ trợ cho nội dung web như STATISTICA, GGobi và T-SNE[1]. Với các công cụ này, người dùng có thể hình dung được các nội dung rộng lớn một cách toàn cảnh dưới các dạng như biểu đồ phân tích thống kê, biểu đồ phân tán, đồ thị, v.v.

2.2. Các kỹ thuật khai thác với dữ liệu có cấu trúc

Để khai thác dữ liệu có cấu trúc, chúng ta sử dụng các kỹ thuật khai thác dữ liệu truyền thống. Dữ liệu có cấu trúc ở các trang web thường có dạng bảng, danh sách và cây. So với dữ liệu phi cấu trúc, việc rút trích dữ liệu có cấu trúc tiến hành dễ dàng hơn, bao gồm những kỹ thuật như sau.

Trình thu thập thông tin web: là một robot công cụ tìm kiếm được thiết kế với mục đích tìm kiếm, thu thập thông tin, là chương trình tự động duyệt qua các URL và tải xuống các trang web tìm thấy có sự liên quan. Đây là một công cụ khá quan trọng trong việc tối ưu website, tối ưu hóa công cụ

[1] STATISTICA: là một bộ phần mềm phân tích thống kê được phát triển cho khai thác dữ liệu, quản lý dữ liệu, trực quan dữ liệu và phân tích dữ liệu.

Ggobi: là một chương trình trực quan mã nguồn mở để khám phá dữ liệu chiều cao.

t-SNE *(t-distributed stochastic neighbor embedding):* phương pháp giảm kích thước để giúp hình dung các đối tượng.

tìm kiếm và giúp website tiếp cận được lượng lớn người dùng truy cập [7].

Tạo trình bao gói: có thể định nghĩa là quá trình xếp hạng các trang web dựa trên giá trị xếp hạng web để truy xuất các trang web có liên quan, bằng các công cụ tìm kiếm theo câu truy vấn do người dùng thực hiện trên trình duyệt web của họ [8].

Khai thác nội dung trang web: là quá trình khai thác toàn bộ nội dung của các trang web bằng các công cụ tìm kiếm. Quá trình này bao gồm việc rút trích dữ liệu có cấu trúc, sau đó liệt kê kết quả dựa trên sự tương quan của chúng [6], [8].

2.3. Các kỹ thuật khai thác với dữ liệu bán cấu trúc

Dữ liệu bán cấu trúc được thu thập kết hợp từ các nguồn không đồng nhất như các trang web trong cơ sở dữ liệu và sách, bao gồm các kỹ thuật sau:

Mô hình trao đổi đối tượng: là quá trình trích xuất thông tin hữu ích từ các nội dung bán cấu trúc, sau đó phân chia các nội dung tương tự nhau thành các nhóm thống nhất [6].

Trích xuất từ trên xuống: là quá trình trích xuất các đối tượng phức tạp từ các nguồn tài nguyên web khác nhau và phân tách chúng dần thành các đối tượng nhỏ nhất có thể trích xuất được và thu thập chúng [10].

Ngôn ngữ trích xuất dữ liệu web: kỹ thuật này được sử dụng để chuyển đổi nội dung web bán cấu trúc thành dạng có cấu trúc [9].

2.4. Các kỹ thuật khai thác với dữ liệu đa phương tiện

Dữ liệu đa phương tiện là loại dữ liệu tổng hợp của nhiều kiểu dữ liệu như văn bản, hình ảnh, âm thanh, audio, và video. Các kỹ thuật khai thác dữ liệu đa phương tiện cũng kế thừa từ kỹ thuật khai thác dữ liệu truyền thống như phân lớp, gom cụm, khai thác mẫu, luật kết hợp và thống kê tiêu chuẩn. Mục tiêu chính của của các hệ thống khai thác dữ liệu đa phương tiện là tìm kiếm độ tương đồng trong dữ liệu đa phương tiện, triển khai xây dựng và truy xuất theo chỉ mục, truy xuất dựa trên mô tả bằng cách sử dụng các *từ khóa, thẻ, tiêu đề, dấu thời gian*; truy xuất dựa trên nội dung như *màu sắc, biểu đồ histogram, kết cấu, hình dạng, biến đổi wavelet.* Khai thác dữ liệu đa phương tiện bao gồm hai bước cơ bản, bao gồm rút trích các đặc trưng từ dữ liệu và chọn lựa phương pháp khai thác để xác định nội dung mong muốn. Khai thác đa phương tiện được phân loại như sau:

Khai thác văn bản (text mining): là quá trình áp dụng các kỹ thuật khai thác dữ liệu để trích xuất các phần (khối) văn bản quan trọng từ các tài liệu web phi cấu trúc. Bag of Words (BOW) là một mô hình phổ biến được sử dụng trong khai thác web để thể hiện sự vắng mặt hoặc hiện diện của các thành phần văn bản (ví dụ như câu hoặc từ), mô hình này còn được gọi là mô hình không gian vectơ (Vector Space Model).

Khai thác ảnh (Image Mining): kỹ thuật này được sử dụng để khám phá các mẫu ảnh hoặc thông tin hữu ích từ tập hình ảnh đồ sộ. Khai thác ảnh gồm một số hướng tiếp cận như xử lý ảnh kỹ thuật số, hiểu nội dung ảnh, trí tuệ nhân tạo, v.v., nhiều phương pháp đề xuất chủ yếu tập trung vào xử lý ảnh để rút trích các đặc trưng mong muốn như phân tích kết cấu, phát hiện đường biên, làm mịn, lập biểu đồ màu, v.v.

Khai thác âm thanh (Audio Mining): là quá trình phân tích và tìm kiếm thông qua nội dung âm thanh, thường được sử dụng trong lĩnh vực nhận dạng giọng nói.

Khai thác video (Video Mining): liên quan đến các nhiệm vụ xử lý video kỹ thuật số bao gồm phân đoạn tự động, lập chỉ mục, truy xuất dựa trên nội dung và phân lớp các đối tượng trực quan.

3. Khai thác cấu trúc web

Khai thác cấu trúc web là sử dụng lý thuyết đồ thị để phân tích và hiểu được cấu trúc kết nối của website [11].

Cấu trúc web có thể được biểu diễn dưới dạng đồ thị trong đó mỗi trang web là một nút của đồ thị và cạnh nối các nút chính là

siêu liên kết kết (hyperlink) nối giữa các trang web. Khai thác cấu trúc web là khám phá thông tin về cấu trúc, mối liên hệ giữa các trang web trong cùng một website hoặc giữa các website. Có thể chia khai thác cấu trúc web thành hai loại tùy thuộc vào loại dữ liệu cấu trúc khai thác: khai thác cấu trúc mức siêu liên kết và khai thác cấu trúc tài liệu web. Siêu liên kết là một đơn vị cấu trúc kết nối từ một vị trí trong một trang web liên kết đến một vị trí khác trong cùng trang hoặc đến trang khác. Còn cấu trúc của một tài liệu web được tổ chức dưới dạng cấu trúc cây, dựa trên các thẻ HTML và XML. Các kỹ thuật khai thác cấu trúc web bao gồm phân cụm, phân lớp và truy tìm thông tin.

Phân cụm: mục đích chính của việc phân cụm là nhóm các trang web tương tự nhau về cấu trúc hyperlink, hoặc cấu trúc tài liệu. Vì vậy, phân cụm giúp người dùng có thể tìm được các thông tin cần thiết thông qua việc kết hợp các từ khóa chứa trong hyperlink hoặc các nội dung rút trích từ cấu trúc tài liệu.

Phân lớp: Đối với khai thác cấu trúc web, có hai loại phân lớp là phân lớp dựa trên nội dung liên kết và phân lớp dựa trên siêu liên kết.

Truy xuất thông tin: thông tin chứa trong các siêu liên kết đóng một vai trò quan trọng để truy xuất kết quả trong các công cụ tìm kiếm, nơi các phần văn bản neo (văn bản xuất hiện trong các siêu liên kết) của các trang Web gốc đã được lập chỉ mục bởi World Wide Web Worms.

4. Khai thác sử dụng web

Phần lớn các trang web có thể được truy cập hàng nghìn lần mỗi ngày, đặc biệt là những trang web thương mại. Vấn đề là làm cách nào để thu thập những thông tin này nhằm phân tích xem người dùng duyệt gì, cần gì để có thể đưa ra những chiến lược quan trọng cho mô hình thương mại của các doanh nghiệp hiện tại. Các thông tin này thường được lưu trữ trong các file nhật ký (web log), dưới dạng file văn bản vì khi người dùng mở các trang web khác nhau thì dấu vết cũng như hành vi duyệt web của người dùng tự động lưu vào file log. Chính vì vậy, khai thác tri thức từ web log sẽ giúp các tổ chức trong việc đưa ra các quyết định kinh doanh, cải tiến, thiết kế trang web đạt đến một đỉnh cao mới trong lĩnh vực thương mại điện tử. Đây là lĩnh vực thu hút nhiều quan tâm nghiên cứu nhất trong ba loại hình khai thác web.

4.1. Qui trình khai thác sử dụng web

Tổng thể quá trình khai thác sử dụng web có thể được chia thành các bước sau: thu thập dữ liệu, tiền xử lý dữ liệu, khai thác mẫu, phân tích đánh giá mẫu.

4.1.1. Thu thập dữ liệu

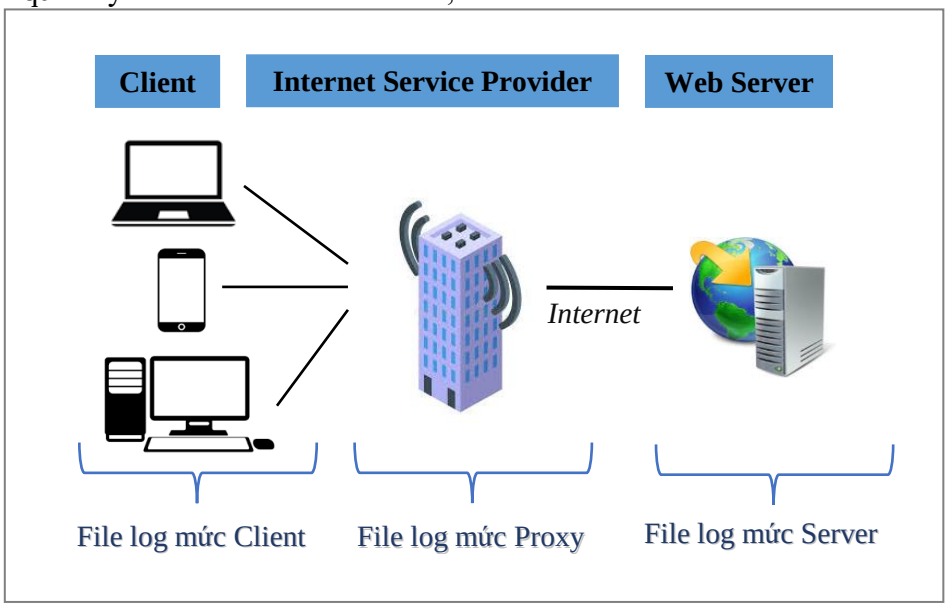
Dữ liệu có thể được thu thập từ các mức khác nhau hoặc lấy từ một tổ chức dưới dạng tập dữ liệu. Dữ liệu được thu thập được có thể khác nhau về cấu trúc nội dung, nguồn, loại thông tin có sẵn, phương pháp phân đoạn và phương pháp triển khai. Các mức nguồn dữ liệu gồm: dữ liệu mức Client (máy khách), dữ liệu mức Proxy, và dữ liệu mức Server (máy chủ) [3]. **Hình 2** biểu diễn kiến trúc các mức nguồn dữ liệu Web.

File log mức Client: được thu thập từ các thiết bị của khách khi truy cập web. Dữ liệu này thu được nhờ các proxy áp đặt từ một số ứng dụng đính kèm trong trang web để lấy thông tin khách truy cập. Tuy nhiên, vì lý do bảo mật của người dùng, dữ liệu này cần có sự đồng ý của khách truy cập.

File log mức Proxy: hoạt động như một khâu trung gian giữa Client và Server, nhận các yêu cầu HTTP từ người dùng và đưa nó đến Web Server, sau đó Web Server trả về kết quả và chuyển lại cho người dùng đã gửi yêu cầu. File log mức Proxy ghi nhận dữ liệu là thông tin máy chủ proxy chứ không phải thông tin của Client ban đầu.

File log mức Server: chứa dữ liệu phía Server, cung cấp dữ liệu sử dụng chính xác và đầy đủ nhất khi người dùng tương tác với các trang Web khác nhau. Đây là nguồn dữ liệu chính trong khai thác sử dụng web,dữ liệu này bao gồm địa chỉ IP, URL, số byte, v.v..

Định dạng của tập tin Web log: mặc dù web log được lưu tại máy chủ nhưng tuỳ theo nhà cung cấp dịch vụ web, máy chủ có thể được cài đặc hệ điều hành, Web server và các phần mềm quản lý web khác nhau. Do đó, cấu trúc của các web log cũng khác nhau, có các định dạng: NCSA common, NCSA combined, định dạng mở rộng W3C format và định dạng IIS.



***Hình 2**. Kiến trúc các mức nguồn dữ liệu web.*

Tập tin web log ghi lại các thông tin cơ bản về các yêu cầu của người dùng đối với một website. Thông tin này được ghi lại dưới dạng: *host/ip user [date:time] "method url" status bytes "ReferenceUrl" "agent",* **Hình 3** minh họa một đoạn thông tin trong file web log gồm 3 lượt truy cập. Trong đó:

- *host/ip*: địa chỉ *host/ip* của máy tính truy cập vào trang web.
- *user*: số định danh người dùng (biểu thị định danh bị giấu đi).
- *[date: time]:* thời gian truy cập.
- *method*: phương thức yêu cầu của người sử dụng web (GET; POST).
- *url*: đường dẫn của trang web được truy cập.
- *status:* tình trạng của yêu cầu (200).
- *byte:* số lượng byte dữ liệu đã yêu cầu.
- *RefernceUrl:* địa chỉ trang web trước mà từ đó dẫn đến địa chỉ hiện tại.
- *agent:* thông tin về hệ điều hành, trình duyệt của máy người sử dụng, chẳng hạn như tên người dùng, ngày, giờ, loại yêu cầu, mã trạng thái HTTP, và số lượng các byte được gửi bởi máy chủ.

Trong số những thông tin được ghi lại trong tập tin log, có những thông tin không cần thiết cho việc khai thác. Do đó, cần có bước tiền xử lý loại bỏ những thông tin dư thừa để thu được CSDL dùng cho việc khai thác mẫu truy cập web **Error! Reference source not found.**.

```
66.249.79.19 - - [15/Jun/2018:04:06:53 +0700] "GET /toyota-corolla-altis-2015.html HTTP/1.0" 200 6794 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
37.140.141.38 - - [15/Jun/2018:04:14:06 +0700] "GET /robots.txt HTTP/1.0" 200 865 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
5.45.254.225 - - [15/Jun/2018:04:14:09 +0700] "GET /san-pham/yaris/Page-2.html HTTP/1.0" 200 6931 "-" "Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)"
```

***Hình 3.** Ví dụ về nội dung của một file web log.*

4.1.2. Tiền xử lý dữ liệu

Trong giai đoạn tiền xử lý, dữ liệu truy cập của người sử dụng web được làm sạch và phân hoạch thành một tập hợp các chuỗi truy cập của người dùng. Tập hợp này đại diện cho các hoạt động của mỗi người dùng trong những lần truy cập trang web.

Khai thác thói quen sử dụng web tiến hành trên dữ liệu được tạo ra bằng cách xem xét các session hoặc các thói quen, chúng được lưu trữ trên web log được rút trích từ các Web server.

Trong thực tế, tiền xử lý dữ liệu web là giai đoạn gặp nhiều thách thức vì dữ liệu web có quy mô khổng lồ, vượt xa bất kỳ loại dữ liệu thông thường nào khác. Hơn nữa, các file log mức Server được ghi lại dưới dạng văn bản có thể ở nhiều định dạng khác nhau. Do đó, quá trình tiền xử lý thường mất nhiều thời gian và chi phí tính toán. Tiền xử lý dữ liệu web gồm các bước làm sạch dữ liệu, xác định người dùng và phiên truy cập.

4.1.3. Khai thác mẫu

Đây là một trong những khâu quan trọng nhất của Khai thác sử dụng web, rút trích dữ liệu từ lịch sử truy cập web. Mục tiêu chính của giai đoạn này là khám phá các mẫu truy cập thú vị từ người dùng.

Thống kê: đây là một kỹ thuật quan trọng để tìm thông tin hữu ích từ bất kỳ lịch sử duyệt web nào, để biết được nội dung lịch sử duyệt web và số lượt truy cập của khách hàng trong bản ghi đó. Số lượt truy cập được tính trên cơ sở mỗi mục nhập hợp lệ trong web log bao gồm đăng bài, duyệt, tải xuống. Kỹ thuật thống kê giúp cải thiện hiệu suất hệ thống, chẳng hạn như theo dõi các hoạt động của khách truy cập, theo dõi và kiểm tra các trang và website cũng như tổng hợp khách truy cập dựa trên hành vi duyệt web của họ [1].

Luật kết hợp: kỹ thuật này được sử dụng để tìm các quy luật từ các mẫu trong dữ liệu được tạo từ giai đoạn tiền xử lý dữ liệu web. Chẳng hạn như tìm số lần truy cập thường xuyên của người dùng vào các trang nhất định. Nhiệm vụ của công nghệ này là hiểu các yêu cầu của khách truy cập, xem thử trong một website khách hàng đã duyệt qua những trang nào và mối quan hệ giữa các trang đó. Đã có nhiều thuật toán đề xuất cho kỹ thuật này điển hình là nhóm các thuật toán Apriori [13].

Mẫu tuần tự: tìm ra các mẫu truy cập phổ biến thông qua các phiên truy cập nối tiếp, bằng cách áp dụng một số thuật toán như SPADE, Apriori, v.v. Phân tích các mẫu truy cập này, chúng ta có thể biết được thói quen cũng như mối quan tâm của khách hàng, từ đó dự đoán nhu cầu mua sắm của khách hàng, đưa ra chiến lược quảng cáo, suy ra những bất thường để khám phá tội phạm, v.v.[12].

Phân cụm: là một phương pháp được sử dụng để nhóm các thành phần nhất định như trang, người dùng… dựa vào đặc điểm giống nhau. Chẳng hạn như nhóm các trang web có nội dung tương tự, nhóm các khách truy cập có hành vi duyệt tương tự, hoặc tập hợp nhiều người dùng truy cập các trang web tương tự nhau. Công nghệ phân cụm giúp suy ra số liệu thống kê về khách hàng trong các hoạt động của thị trường thương mại điện tử. Và cung cấp nội dung web tùy chỉnh dựa trên khách truy cập cá nhân. Ngoài ra, phân cụm rất hữu ích trong việc tạo chỉ mục của các trang web trên Internet hỗ trợ cho việc tìm kiếm [12].

Phân lớp: có nhiều thuật toán có thể được sử dụng để phân loại người dùng vào tập các lớp đã được xác định trước, quá trình này bao gồm xây dựng thuật toán phân loại nhờ dữ liệu huấn luyện sau đó được sử dụng thuật toán đích để phân loại dữ liệu thực nghiệm [5].

4.1.4. Phân tích đánh giá mẫu

Đây là giai đoạn cuối cùng của khai thác sử dụng web nhằm rút trích những tri thức từ tập mẫu thú vị khai thác được, loại bỏ các mẫu không phù hợp. Trong giai đoạn cuối của quá trình, các mẫu phát hiện và thống kê được tiếp tục xử lý, lọc và có thể được sử dụng làm đầu vào cho các ứng dụng như công cụ trực quan, phân tích web và các công cụ tạo báo cáo. Phân tích mô hình, thống kê,

tìm kiếm tri thức và tác nhân thông minh. Phân tích tính khả thi, truy vấn dữ liệu hướng tới sự tiêu dùng của con người.

4.2. Ứng dụng của khai thác sử dụng web

Mục tiêu chung của khai thác sử dụng web là tích lũy các dữ liệu hấp dẫn về thiết kế tuyến truy cập của khách hàng. Dữ liệu này có thể được sử dụng về sau để cải thiện website theo góc nhìn của khách hàng. Kết quả thu được từ việc khai thác web log có thể được sử dụng cho các mục đích khác nhau như để cá nhân hóa việc cung cấp nội dung trang web; để cải thiện điều hướng người dùng thông qua tìm nạp trước và lưu bộ nhớ đệm; để cải thiện thiết kế web hoặc các trang thương mại điện tử; và để cải thiện sự hài lòng của khách hàng.

5. Kết luận

Khai thác web trở thành một lĩnh vực không thể thiếu đối với nhiều ứng dụng phát triển web trong khi nội dung thông tin dữ liệu của các cá nhân, tổ chức, và doanh nghiệp trên web ngày càng gia tăng. Trong bài báo này, chúng tôi khảo sát các kỹ thuật khai thác web khác nhau được sử dụng bởi rất nhiều ứng dụng web gần đây. Chúng tôi cũng đã xem xét và so sánh giữa các loại khai thác web dựa trên các phương pháp tiếp cận quan trọng được hầu hết các công trình nghiên cứu sử dụng hiện nay.

Vì môi trường web hàm chứa phạm vi nghiên cứu rộng lớn và có rất nhiều tác vụ cần nghiên cứu trong tương lai, chúng tôi hy vọng bài báo này có thể cung cấp nền tảng khởi đầu về bản chất dữ liệu web cũng như các kỹ thuật khai thác dữ liệu hiện đang được áp dụng trên các dữ liệu web khác nhau; từ đó giúp định hướng các cơ hội nghiên cứu trong từng lĩnh vực khai thác web cụ thể.

Tài liệu tham khảo

[1] Shyam N. Kumar (2015), *"World towards Advance Web Mining: A Review"*, American Journal of Systems and Software, Vol.3, No.8, pp. 44-61.

[2] Faustina Johnson, Santosh Kumar Gupta (2012), *"Web Content Mining Techniques: A Survey"*, International Journal of Computer Applications, Vol.47, No.11, pp. 44-50.

[3] Monika Yadav, Pradeep Mittal (2013), *"Web Mining: An Introduction"*, International Journal of Computer Science and Software Engineering, Vol.3, No.3, pp. 683-688.

[4] Al-asadi, T. A., Obaid, A. J., Hidayat, R., & Ramli, A. A. (2017), *"A survey on web mining techniques and applications"*, International Journal on Advanced Science Engineering and Information Technology, 7(4), 1178-1184..

[5] Anurag kumar, Ravi Kumar Singh (2017), *"A Study on Web Content Mining"*, International Journal of Engineering and Computer Science, Vol.6, No.1, pp: 20003-20006.

[6] Athena Vakali, George Pallis, Lefteris Angelis (2007), *"Clustering Web Information Sources"*; In Web Data management practices: Emerging Techniques and Technologies, IDEA group publishing, pp. 34-55.

[7] Filippo Menczer (2011), "*Web Crawling*", in Web Data Mining, Exploring Hyperlinks, Contents and Usage data, Bing Liu, 2nd edition, Springer New York, pp: 311-362.

[8] Sarla More, Durgesh K. Mishra (2012), *"Multimedia Data Mining: A Survey"*, PRATIBHA, International Journal of Science, Spirituality, Business and Technology, Vol.1, No.1, pp. 49-55.

[9] Leslie F. Sikos (2015), *"Mastering Structured Data on the Semantic Web from HTML5 Microdata to Linked Open Data"*, Apress, New York, , pp: 256.

[10] Hashemi, M (2020), *"Web page classification: a survey of perspectives, gaps, and future directions"*, Multimed Tools Appl 79, 11921-11945.

[11] da Costa, M. G., & Gong, Z. (2005, June), *"Web structure mining: an introduction"*, In 2005 IEEE International Conference on Information Acquisition (pp. 6-pp). IEEE.

[12] Neha Sharma (2017), *"A Review on Analysis to Improve Performance of Website"*, International Journal of Science and Research, Vol.6, No.6, pp. 2453-2455.

[13] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta (2013), *"Web Usage Mining using Association Rule Mining on Clustering Data for Pattern Discovery"*, International Journal of Data Mining techniques and Application, Vol.2, No.1, pp. 141-150.