PHÂN TÍCH TẬP TIN NHẬT KÝ SỬ DỤNG KỸ THUẬT KHAI PHÁ VÀ LOGIC MỜ

Nguyễn Văn Quân[1*], Hoàng Tuấn Hảo[1], Vũ Văn Cảnh[1], Hoàng Thế Triều[2]

Tóm tắt: *Cùng với lượng dữ liệu Website ngày càng tăng nhanh trên Internet, trong những năm gần đây lĩnh vực nghiên cứu khai phá dữ liệu Website rất được quan tâm. Trong bài báo này, chúng tôi nghiên cứu khái quát một số kỹ thuật khai phá và logic mờ nhằm khai phá theo sử dụng Website dựa trên phân tích tập tin nhật ký - ghi lại hoạt động của người dùng khi tương tác với Website. Trong quá trình nghiên cứu cũng thực hiện kỹ thuật phân cụm mờ và kết hợp luật mờ nhằm nâng cao hiệu quả kiểm tra tập dữ liệu nhật ký từ máy chủ Webserver.*

Từ khóa: Khai phá Web, Logic mờ, Tập tin nhật ký, Fuzzy.

1. GIỚI THIỆU

Trong những năm gần đây cùng với sự phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ về tri thức. Kho dữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vấn đề khai thác các nguồn tri thức đó đặt ra thách thức lớn cho ngành công nghệ thông tin của thế giới.

Cùng với sự tiến bộ vượt bậc của ngành công nghệ thông tin và sự phát triển mạnh mẽ của mạng thông tin toàn cầu, nguồn dữ liệu Web đã trở thành kho dữ liệu khổng lồ. Số lượng Website tăng mạnh, dữ liệu Website vô cùng lớn đòi hỏi phát triển nhiều kỹ thuật quản lý, lưu trữ và khám phá tri thức trên cơ sở dữ liệu lớn – Knowledge Discovery in Database (KDD). Giai đoạn chính của KDD là quá trình khai phá dữ liệu, thông qua kỹ thuật khám phá thì tri thức có thể được tìm thấy trong dữ liệu, và nó thường được lưu trữ trong cơ sở dữ liệu quan hệ theo một dạng cấu trúc [1]. Các lĩnh vực nghiên cứu khác cũng phát triển liên quan tới Web và khai thác thông tin tài liệu trong cơ quan và tổ chức. Công nghệ Web thay đổi, phát triển nhanh chóng và ngày càng được mở rộng không đơn thuần chỉ để tìm kiếm và truy vết thông tin mà còn để thiết lập các giao dịch thương mại. Sự cạnh tranh trong thương mại điện tử đưa ra yêu cầu tạo các ứng dụng thông minh để lưu trữ, khảo sát thông tin về các phiên sử dựng Web hoặc thông tin về khách hàng tiềm năng. Chính vì lý do này, hành vi và đối tượng người dùng là yếu tố cần thu thập và phân tích. Cơ sở dữ liệu tri thức về người dùng được sử dụng không chỉ để mô tả về người dùng mà còn để khám phá các khuynh hướng chung phục vụ cho mục đích thương mại và để cải thiện chất lượng của chính các Website. Dữ liệu tri thức về người dùng được thu thập, lựa chọn từ hành vi của người dùng trong quá trình truy cập Website thông qua các tập tin nhật ký.

Mục tiêu khai phá tập tin nhật ký trong Webserver nhằm xác định mối quan hệ giữa người dùng và những khía cạnh khác có liên quan. Tính chất tự nhiên của dữ liệu tri thức trong các tập tin nhật ký và thông tin để dự đoán như thời gian, tuổi người dùng, trình độ văn hóa…thường được thực hiện bằng kỹ thuật logic mờ. Đây là một công cụ được sử dụng để mô hình hóa thông tin liên quan đến khai phá Web.

Trong bài báo này, chúng tôi trình bày tóm tắt một số nghiên cứu sử dụng logic mờ trong khai phá dữ liệu Web. Với mục đích giải thích ba dạng khai phá dữ liệu Web: Khai phá nội dung Web, khai phá cấu trúc Web và khai phá theo sử dụng Web. Sau đó tập trung vào khai phá theo sử dụng Web bao gồm nghiên cứu các quá trình cá nhân hóa và xây dựng hồ sơ người dùng trên Web. Chúng tôi tóm lược các ứng dụng chính của logic mờ trong một số công trình nghiên cứu và mô tả một số thí nghiệm sử dụng logic mờ trong khai phá dữ liệu Web.

2. KHAI PHÁ WEB

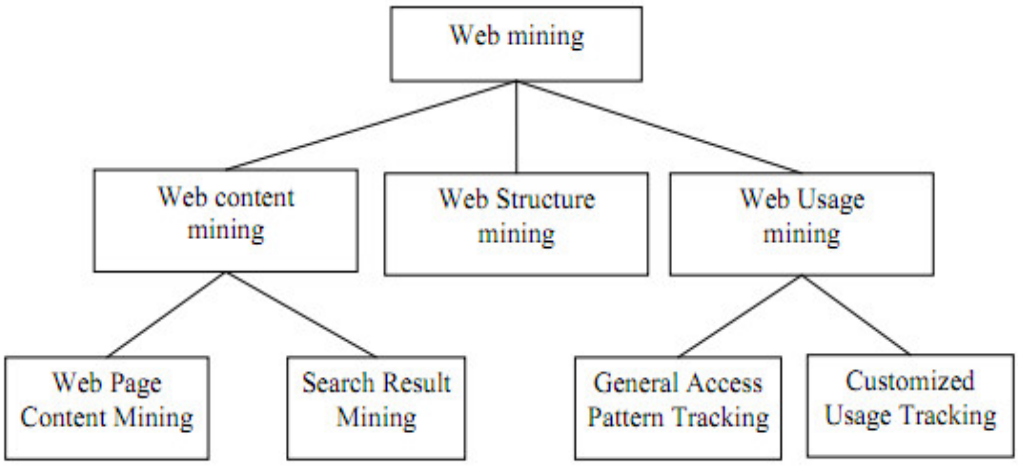
Có nhiều khái niệm khác nhau về khai phá Web, nhưng có thể tổng quát hóa như sau [16]: Khai phá Web là việc sử dụng các kỹ thuật khai phá dữ liệu để tự động hóa quá trình khám phá và trích rút những thông tin hữu ích từ các tài liệu, các dịch vụ và cấu trúc Web. Nói cách khác khai phá Web là quá trình thăm dò những thông tin quan trọng, các mẫu tiềm năng từ nội dung Web, từ thông tin truy cập Web, từ liên kết trang và từ nguồn tài nguyên thương mại điện tử bằng các kỹ thuật khai phá dữ liệu, giúp con người trích rút các tri thức, cải tiến quá trình thiết kế Website và phát triển tốt hơn trong lĩnh vực thương mại điện tử.

Những thách thức gặp phải trong quá trình thu thập thông tin cần thiết: Số lượng dữ liệu lớn, ngôn ngữ đa dạng, vấn đề chất lượng thông tin, sự phân bố dữ liệu trên các nền tảng khác nhau và cuối cùng rất quan trọng đó là sự thiếu cấu trúc trong dữ liệu Web. Từ những đặc điểm trên, đặc biệt, đối với dữ liệu phi cấu trúc và tính không đồng nhất cũng là những điểm khó khăn chính của quá trình khai phá Web. Trong những quá trình này, các kỹ thuật khai phá dữ liệu được sử dụng để khám phá tự động và trích chọn thông tin từ các tài liệu và các dịch vụ Web [12].

Cooley đưa ra ba hình thức khai phá Web: Xuất phát từ nội dung, cấu trúc và theo sử dụng [6].

Khai phá nội dung Web là khám phá tự động các mẫu từ nội dung văn bản Web [7][21]. Khai phá cấu trúc Web bao gồm nghiên cứu về cấu trúc liên kết đưa vào hoặc nội dung các văn bản bên trong để khám phá các mẫu hữu ích của cấu

trúc liên kết [7][9]. Cuối cùng là khai phá theo sử dụng Web, đây là nội dung chính chúng tôi sẽ đề cập trong nghiên cứu này. Chúng tôi có thể định nghĩa đây là tiến trình khám phá tự động mẫu truy cập hoặc sử dụng các dịch vụ Web, dựa trên hành vi người dùng khi tương tác với Web [10]. Chúng tôi sẽ tập trung thảo luận về khai phá theo sử dụng Web trong phần tiếp theo.



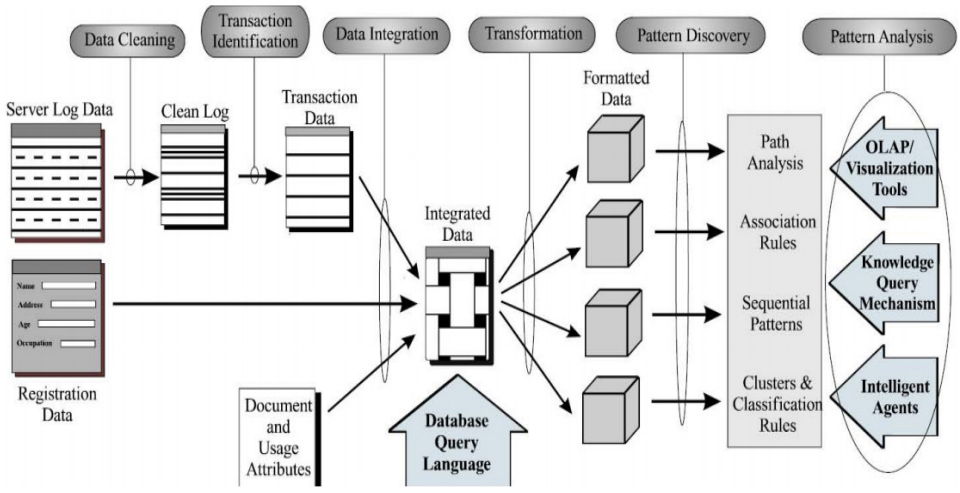
Hình 1. *Phân loại khai phá Web.*

2.1. Khai phá theo sử dụng Web

Việc thu thập các thông tin về người dùng có ý nghĩa rất quan trọng đối với người thiết kế Website. Thông qua quá trình khai phá lịch sử các mẫu truy cập của người dùng Web, không chỉ thông tin về Web được sử dụng như thế nào mà còn nhiều đặc tính khác như các hành vi của người dùng có thể được xác định. Sự điều hướng đường dẫn người dùng Web mang lại giá trị thông tin về mức độ quan tâm của người dùng đối với Web. Dựa trên các tiêu chuẩn khác nhau thì người dùng Web có thể được phân cụm và các tri thức hữu ích có thể được lấy ra từ các mẫu truy cập Web. Nhiều ứng dụng có thể giúp lấy ra được các tri thức. Thông qua việc phát hiện mối quan hệ giữa những người dùng có cùng sở thích, sự quan tâm của người dùng Web ta có thể dự đoán chính xác hơn về người dùng đang cần gì, tại thời điểm hiện tại có thể dự đoán kế tiếp theo họ sẽ truy cập các thông tin gì.

Khi người dùng tương tác với Website, họ để lại thông tin dấu vết dạng số (IP, agent, cookies...) được server tự động lưu trữ trong nhật ký truy cập. Các tập tin nhật ký chứa thông tin kết nối máy chủ hoặc định danh người dùng và xác thực. Những thông tin này thu thập hành vi người dùng trên mạng và phản ánh một số loại mô hình khác nhau về hành vi.

Khai phá sử dụng Web là khai phá truy cập Web (Web log) để khám phá các mẫu người dùng truy cập vào Website. Qua quá trình phân tích và khảo sát các quy tắc trong việc ghi nhận lại quá trình truy cập Web ta có thể chứng thực khách hàng trong thương mại điện tử, nâng cao chất lượng dịch vụ thông tin trên Internet đến người dùng, nâng cao hiệu suất của các hệ thống phục vụ Web. Thêm nữa, phân tích quá trình đăng nhập Web của người dùng giúp cho việc xây dựng các dịch vụ Web theo yêu cầu đối với từng người dùng sẽ tốt hơn. Hiện nay, ta thường sử dụng các công cụ khám phá mẫu và phân tích mẫu. Nó phân tích các hành động người dùng, lọc dữ liệu và khai phá tri thức từ tập tin dữ liệu bằng cách sử dụng trí tuệ nhân tạo, khai phá dữ liệu, tâm lý học và lý thuyết thông tin. Kiến trúc tổng quát của quá trình khai phá theo sử dụng Web như sau:



Hình 2. *Kiến trúc tổng quát của quá trình khai phá theo sử dụng Web.*

2.2. Những vấn đề trong khai phá theo sử dụng Web

Khai phá theo sử dụng Web có hai quá trình cần thực hiện: Thứ nhất là Web-log cần được làm sạch, định nghĩa, tích hợp và biến đổi; Thứ hai là phân tích và khai phá. Có nhiều vấn đề khó khăn nảy sinh ở đây như cấu trúc vật lý của các Website khác nhau từ những mẫu người dùng truy xuất hoặc rất khó để có thể tìm ra những phiên người dùng, các phiên làm việc và các thao tác.

Khả năng của Website để xử lý một tương tác với mức độ chi tiết và hướng dẫn khách hàng hoặc người dùng thông qua các thông tin hữu ích và cần thiết thành công, đang trở thành một trong những mục tiêu quan trọng cho mọi Website ngày nay. Một trong những giải pháp để đạt được mục tiêu này là thông qua sự cá nhân hóa của Website.

Sự cá nhân hóa Website có thể được nhìn nhận từ hai quan điểm: Phía công ty và phía người dùng. Quan điểm của công ty là mối quan hệ giữa tiếp thị và định danh lớp khách hàng. Quan điểm người dùng là mối quan hệ giữa sự khuyến nghị và thu thập thông tin. Quá trình này có thể mô tả như nhóm các hành vi được thực hiện bởi người dùng, những hành động này có thể được xử lý để cải thiện Website theo sở thích của người dùng [22]. Phần thông tin này có thể được lưu trong hồ sơ người dùng. Hồ sơ người dùng có thể được định nghĩa như biểu diễn tri thức về thông tin sở thích của người dùng [20], các tác giả đề xuất hai dạng hồ sơ khác nhau: Hồ sơ đơn thuần được biểu diễn bởi trích xuất dữ liệu từ tài liệu được cho là người dùng quan tâm; và các hồ sơ mở rộng có chứa các tri thức bổ sung về người dùng như tuổi, trình độ ngoại ngữ, quốc tích và một số thông tin bổ sung khác.

Đối với việc thu thập những hồ sơ này thì sự phân cụm và các quy tắc kết hợp thường được áp dụng. Thông qua quá trình phân cụm, một nhóm khách hàng hoặc dữ liệu với các đặc tính tương tự được tự động khởi tạo thu thập mà không có sự phân loại trước đây. Hồ sơ người dùng bắt nguồn từ các nhóm này có thể được sử dụng để chỉ dẫn các chiến lược tiếp thị theo nhóm [23]. Các quy tắc kết hợp khám phá sự kết hợp và sự tương quan giữa các mặt hàng trong đó sự hiện diện của một mặt hàng hoặc một nhóm trong giao dịch ngụ ý rằng có sự có mặt của các mặt hàng khác [4]. Một ứng dụng trực tiếp nhất của quy tắc kết hợp để khai phá người dùng web xuất phát từ mối quan hệ giữa sự ghé thăm của người dùng với mô hình định hướng nhất định cho website.

Sự bất tiện chính của các hồ sơ trong Website là thiếu tri thức về danh tính của người dùng. Hai tình huống khác nhau có thể gia tăng: Thứ nhất, người dùng chưa đăng ký trong đó hồ sơ người dùng có thể cung cấp bằng chứng về danh tính hoặc liên kết với một nhóm mạng xã hội. Một hồ sơ chung sau đó được ấn định cho người dùng. Sự tùy chọn lưu trữ trong hồ sơ có thể được áp dụng cho Website để người dùng đăng ký.

Tình huống thứ hai người dùng đã đăng ký, nếu một người dùng được nhận diện theo phương pháp nào đó, Website có thể thay đổi tùy theo sở thích của người dùng. Hệ thống sẽ lưu trữ lại dấu vết của người dùng trong những lần ghé thăm trước với hồ sơ người dùng. Để mô tả đặc điểm nhóm người dùng với hành vi tương đồng, có thể thực hiện theo phương pháp phân cụm [20].

Các hành động được thực hiện bởi người dùng từ khi bắt đầu truy cập vào Web cho đến khi rời khỏi Web được ghi nhận và lưu trữ trong một tập tin nhật ký (logfile). Tập tin nhật ký sẽ chứa địa chỉ IP của máy khách, ngày, thời gian từ khi

yêu cầu được tiếp nhận, các đối tượng yêu cầu và các thông tin trong phiên làm việc của người dùng, ví dụ:

```
216.239.46.60 - - [04/April/2007:14:56:50 +0200] "GET
/~lpis/curriculum/C+Unix/ Ergastiria/Week-7/filetype.c.txt HTTP/1.0" 304 -
216.239.46.100- - [04/April/2007:14:57:33 +0200]"GET /~oswinds/top.html
HTTP/ 1.0" 200 869
64.68.82.70 - - [04/April/2007:14:58:25 +0200] "GET /~lpis/systems/rdevice/r-
device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/April/2007:14:58:27 +0200] "GET /~lpis/publications/crc-
chapter1. html HTTP/1.0" 304 -
209.237.238.161 - - [04/April/2007:14:59:11+0200] "GET /robots.txt HTTP/1.0"
404 276
209.237.238.161 - - [04/April/2007:14:59:12 +0200] "GET /teachers/pitas1.html
HTTP/1.0" 404 286
216.239.46.43 - - [04/April/2007:14:59:45 +0200] "GET /~oswinds/publication
```

***Hình 3**. Minh họa nội dung tập tin nhật ký.*

2.3. Một số công trình trước đây

Trong [22] đã khái quát quá trình cá nhân hóa dựa trên khai phá người dùng Website, các kỹ thuật khai phá dữ liệu như phân cụm để khám phá các nhóm người dùng được sử dụng. Hơn nữa, các quy tắc kết hợp có thể được sử dụng để tìm các mối quan hệ quan trọng giữa mục người dùng quan tâm dựa trên các mẫu thông tin chỉ dẫn. Một đề xuất khác về phương pháp thang phân cụm lấy ý tưởng từ hệ thống miễn dịch học tự nhiên cho phép học liên tục và tự đáp ứng với các mẫu mới [20].

WebMiner, một hệ thống nổi tiếng được phát triển cho sự cá nhân hóa dựa trên mô hình hành vi điều hướng của người dùng [6]. Bằng cách nhóm các Website tham khảo, hệ thống tạo ra các giao dịch từ các quy tắc kết hợp được khám phá. Một hệ thống liên quan khác để cá nhân hóa được biểu diễn trong [5], các tập tin nhật ký trên máy chủ được lưu trữ và phân tích. Từ các giao dịch, các mẫu hành vi được trích xuất để mô tả phương thức người dùng lướt web theo phương pháp phân cụm và các quy tắc kết hợp. Trong [24], các tác giả đề xuất một cấu trúc hướng dẫn cá nhân hóa và đáp ứng trong Website bởi hồ sơ người dùng và các truy cập được lựa chọn thông qua các tập tin nhật ký Website.

Tiếp theo, chúng tôi dẫn giải một số đề xuất thực hiện trong lĩnh vực này được kết nối với logic mờ.

3. KHAI PHÁ WEB VỚI LOGIC MỜ

Cũng giống như trong khai phá dữ liệu truyền thống, xét từ góc độ dữ liệu hoặc kỹ thuật thì các công cụ tối ưu nhằm khai phá Web được xây dựng từ tính toán

mềm đã được nghiên cứu và áp dụng như logic mờ, giải thuật di truyền, mạng nơ ron nhân tạo hoặc tập thô [2][15]. Trong khai phá Web, logic mờ có thể trợ giúp việc biểu diễn người dùng lựa chọn theo định hướng dữ liệu, nâng cao sự linh hoạt của hệ thống và tạo ra các giải pháp rõ ràng hơn [21].

Gần đây, các kỹ thuật này được áp dụng vào nhiều lĩnh vực khai phá dữ liệu khác nhau như lựa chọn tài liệu [26] và khai phá Web. Trong khai phá Web, các kỹ thuật thường được sử dụng như phân cụm mờ và các luật kết hợp mờ. Các kỹ thuật này được sử dụng để tìm khuynh hướng chỉ dẫn chung của người dùng và xây dựng hồ sơ người dùng.

Các thuật toán phân cụm mờ như FCM (Fuzzy C-Means), FCTM (Fuzzy-C Trimmed Medoids), và FCLMedS (Fuzzy-C Medians) được sử dụng để khai phá nội dung và người dùng website [21]. Một ứng dụng khác với phân cụm mờ được sử dụng để khai phá cấu trúc và người dùng website [23]. Các tác giả áp dụng thuật toán "*tích tụ cạnh tranh trên các dữ liệu quan hệ*" (CARD - Competitive Agglomeration of Relational Data) để nhóm các phiên người dùng khác nhau. Với mục đích này, không chỉ các mục trong tập tin nhật ký được xem xét mà tính toán sự giống nhau giữa hai phiên người dùng. Mục tiêu của ứng dụng này nhằm xác định phiên người dùng từ các truy cập người dùng vào các Website và cấu trúc của nó.

Cùng với phân cụm mờ, một trong những kỹ thuật ngày càng được sử dụng trong khai phá Website là các luật kết hợp mờ. Một ứng dụng của kỹ thuật này được đề xuất trong [13], trong đó, sự sàng lọc các truy vấn từ một nhóm khởi tạo tài liệu dấu vết lấy từ Website được thực hiện. Các văn bản giao dịch được xây dựng cùng với giá trị mờ. Mục đích của công việc này là cung cấp cho hệ thống khả năng tái lập các truy vấn sử dụng công nghệ khai phá.

Một cách tiếp cận khác sử dụng luật kết hợp mờ, trong [24], tác giả đề xuất kiến trúc hệ thống dự đoán truy cập Website. Các luật kết hợp và thế hệ cây chỉ mục mờ được sử dụng để cải thiện độ chính xác và hiệu suất dự báo trên đường dẫn truy cập Website.

3.1. Logic mờ và hồ sơ người dùng

Logic mờ được phát triển từ lý thuyết tập mờ để lập luận xấp xỉ thay vì lập luận chính xác theo logic vị từ cổ điển [25]. Nó cho phép thao tác và khai thác dữ liệu không đầy đủ hoặc không chắc chắn, đây là điều thường xuyên trong xảy ra trong khai phá dữ liệu [10]. Logic mờ cho phép độ liên thuộc có giá trị trong khoảng đóng [0,1] và ở hình thức ngôn từ, các khái niệm không chính xác như "*hơi hơi*", "*gần như*", "*khá là*", "*rất*". Cụ thể nó cho phép quan hệ thành viên không đầy đủ

giữa thành viên và tập hợp. Lý thuyết này liên quan đến tập mờ và lý thuyết xác suất.

Trong quá trình khai phá sử dụng Web, đôi khi chúng ta không có thông tin chính xác của người dùng trong các tập tin nhật ký ngoài những thông tin nhận được từ server. Để nhận được các thông tin chính xác của người dùng, chúng ta có thể bổ sung thêm định danh của người dùng và xác thực thông qua nguồn dữ liệu khác hoặc có thể suy luận từ các thông tin trong quá trình khai phá. Ví dụ, chúng ta có thể suy luận từ trình độ văn hóa của người dùng dựa vào thói quen của người dùng hoặc từ các thông tin liên quan đến trình độ văn hóa.

Vì vậy, khi hồ sơ người dùng mở rộng được xây dựng, có những thông tin liên quan đến các khái niệm khác nhau về người dùng. Một số khái niệm như độ tuổi của người dùng không chính xác, vì hệ thống phải ước lượng các dữ liệu nếu người dùng không tương xứng, hoặc kiên nhẫn chờ đợi người dùng khai báo trên Website. Các đặc điểm này có thể được mô hình hóa bằng các nhãn ngôn ngữ [20].

Chúng ta thấy các khía cạnh khác nhau cũng như các giải pháp được đề xuất trong lĩnh vực khai phá sử dụng web, chủ yếu dựa trên luật kết hợp và kỹ thuật phân cụm. Nghiên cứu của chúng tôi dựa trên các kỹ thuật này cùng với logic mờ sẽ thu được kết quả có ý nghĩa hơn. Vì thế, luật kết hợp mờ cho phép chúng tôi tìm ra các luật có liên quan đến hành vi người dùng. Trong phần tiếp theo chúng tôi sẽ giải thích về luật kết hợp mờ và thử nghiệm các kỹ thuật có liên quan.

3.2. Luật kết hợp mờ

Luật kết hợp được giới thiệu từ năm 1993, bài toán khai phá luật kết hợp nhận được rất nhiều quan tâm của nhiều nhà khoa học. Ngày nay, việc khai phá các luật như thế vẫn là một lĩnh vực quan trọng trong khai phá dữ liệu. Luật kết hợp giúp chúng ta tìm được các mối liên quan giữa các mục dữ liệu (items) của cơ sở dữ liệu (CSDL) [1]. Luật kết hợp là dạng khá đơn giản nhưng mang lại nhiều hiệu quả. Thông tin về các dạng luật này rất quan trọng và hỗ trợ không nhỏ trong quá trình ra quyết định.

Các luật kết hợp mờ thường tìm kiếm các mối quan hệ hay sự tương đồng giữa các nhóm hạng mục hoặc các lĩnh vực trong một cơ sở dữ liệu quan hệ. Cho I là tập các phần tử được gọi là "*Items*" và cho T là tập các phần tử "*giao dịch*", mỗi giao dịch là một tập các *Items*. Hãy xem xét hai tập *Items* $I_1, I_2 \subseteq I$, trong đó $I_1 \cap I_2 = \varnothing$. Một luật kết hợp $I_1 \Rightarrow I_2$ chỉ sự xuất hiện của các tập phổ biến I_1 trong giao dịch tạo sẽ ra sự xuất hiện của I_2 trong cùng một giao dịch, tuy nhiên, không nhất thiết cần phải có sự đối ứng [17]. I_1 và I_2 được gọi là nguyên nhân và

kết quả của các luật tương ứng. Các biện pháp được dùng để mô tả mối quan hệ giữa nguyên nhân và kết quả của luật kết hợp là "*độ hỗ trợ*", và "*độ tin cậy*". Độ hỗ trợ là tỷ lệ với các giao dịch trong các luật và độ tin cậy đo lường độ chính xác của các luật hay là tỷ lệ của I_1 trong giao dịch có thể tạo ra I_2 trong giao dịch đó.

Một số tác giả đã đề xuất các luật kết hợp mờ để giải quyết các bài toán với dữ liệu mờ hoặc đã được mờ hóa [3][10][14][18][19], các luật kết hợp mờ có thể được trích xuất từ nhóm các giao dịch mờ sử dụng thuật toán APrioriTID [1].

Một giao dịch mờ có thể được định nghĩa là một tập con khác rỗng $\tilde{\tau} \subseteq I$, với mỗi $i \in I$ thì $\tilde{\tau}(i)$ là bậc thành viên i trong giao dịch mờ $\tilde{\tau}$ [12]. $\tilde{\tau}(I_0)$ với $I_0 \subseteq I$ là mức độ hòa nhập của *Item* trong một giao dịch mờ $\tilde{\tau}$, được định nghĩa trong công thức (1):

$$\tilde{\tau}(I_0) = \min_{i \in I} \tilde{\tau}(i) \tag{1}$$

Do đó, các giao dịch mờ điều khiển tính không minh bạch và tạo ra sự linh hoạt hơn, bởi vì chúng cho phép xử lý các giá trị trung gian trong khoảng [0,1] để biểu diễn bậc thành viên của *Items* trong giao dịch.

Để đánh giá việc thực hiện các luật kết hợp, chúng tôi sử dụng theo cách tiếp cận ngữ nghĩa dựa trên việc đánh giá câu định lượng [25]. Một câu định lượng là một biểu thức có dạng "Q của F là G", trong đó, F và G hai tập con mờ của tập hữu hạn X, và Q là lượng hóa mờ tương đối. Định lượng tương đối là các nhãn ngôn ngữ có thể được biểu diễn bằng các giá trị mờ trong khoảng [0,1], chẳng hạn như các nhãn "*hầu hết*", "*hầu như*", hoặc "*nhiều*". Bằng phương pháp này, chúng tôi có thể xác định được ước lượng các luật. Do đó, độ tin cậy và độ hỗ trợ (tỷ lệ xuất hiện) đạt được phụ thuộc vào phương pháp đánh giá và sự lựa chọn lượng hóa. Chúng tôi đánh giá các câu định lượng theo phương pháp GD [8]. Phương pháp này đã được minh chứng đạt được hiệu suất cao hơn các phương pháp đề xuất khác. Công thức để đánh giá "Q của F là G" theo phương pháp GD được định nghĩa trong (2):

$$G\mathrm{D}_Q\left(\frac{G}{F}\right) = \sum_{\alpha_i \Delta\left(\frac{G}{F}\right)} (\alpha_i - \alpha_{i+1}) Q\left(\frac{\left|(G \cup F)_{\alpha_i}\right|}{F_{\alpha_i}}\right) \tag{2}$$

Yếu tố chắc chắn của một luật kết hợp mờ có giá trị trong khoảng [0,1] [8]; Cho một dẫn xuất luật A → C, khi đó yếu tố chắc chắn là tích cực chỉ khi sự phụ thuộc giữa A và C là tích cực, trường hợp giữa A và C độc lập nhau thì yếu tố

chắc chắn là 0, trong trường hợp A và C là đối nghịch thì nó mang giá trị âm. Chúng tôi cho rằng một luật kết hợp mờ là mạnh khi yếu tố chắc chắn của nó và sự hỗ trợ lớn hơn hai giá trị ngưỡng do người dùng định nghĩa tương ứng là "*độ tin cậy nhỏ nhất*" (minCF) và "*sự hỗ trợ/tỷ lệ xuất hiện bé nhất*" (minSupp).

3.3. Thử nghiệm và đánh giá

Trong quá trình thử nghiệm, chúng tôi đã xem xét nhiều kỹ thuật liên quan đến khai phá sử dụng Web, khi tiến hành thực nghiệm chúng tôi áp dụng mô hình tìm kiếm thông tin qua các luật kết hợp mờ. Chúng tôi sử dụng dữ liệu để phân tích từ bộ dữ liệu tập tin nhật ký được đề xuất trong hội nghị ECML/PKDD năm 2005 [11], các tập tin có định dạng CSV. Trong bảng 1 biểu diễn một dòng trong tập tin nhật ký, trong đó bao gồm 6 trường (ID Shop, Date, IP, Session, Visited page, Referenced page).

Bảng 1. *Biểu diễn thông tin một dòng trong tập tin nhật ký.*

ID Shop	**Date**	**IP**
11	Tue Jan 20 19:00:132004	213.235.141.105
Session	**Visited page**	**Referenced Page**
1f75ccd2afbf87dc9abccde23f3	/dt/?c=11670	http://www.shop2.cz/ls/index.php

Mỗi lần thực hiện phân tích một giao dịch, chúng tôi có thể quyết định được dạng thông tin có thể đạt được dựa trên các trường được chọn để tham gia vào các luật thực hiện huấn luyện. Nếu người dùng chọn trường ngày và trang truy cập, các tri thức trích xuất có thể cung cấp kết quả về những trang đã được truy cập nhiều trong một thời gian nhất định (giờ). Ngoài ra, nếu người sử dụng chọn các trường địa chỉ IP và các trang truy cập, chúng ta có thể xác định lượng người dùng đã truy cập vào trang có địa chỉ này. Để nhận được mọi thông tin từ tập tin nhật ký Web, chúng tôi sử dụng thuật toán AprioriTID [1] và các luật kết hợp để trích xuất nhằm giảm số nhóm cần được xem xét. Kết quả chúng tôi có thể nhận được để biết các Website mà người dùng truy cập bắt đầu từ một trang được truy cập ban đầu.

Hình thức các quy tắc được sử dụng để trích xuất là:

Trang khởi tạo ban đầu → Trang tham chiếu

1. dt/?c=11670 → *http://www.shop2.cz/ls/index.php*

- Hỗ trợ (Support) = *0.6*

- Sự tin cậy (Confidence) = *1.0*

- Yếu tố chắc chắn = 1.0

2. dt/?c=12397 → http://www.shop7.cz/akce/kat=239

- Hỗ trợ (Support) = 0.2

- Sự tin cậy (Confidence) = 1.0

- Yếu tố chắc chắn = 1.0

Hai luật được trích xuất từ một tập nhỏ các giao dịch trong đó luật 1 xuất hiện với tỷ lệ 60% và luật 2 xuất hiện với tỷ lệ 20%. Trong cả hai trường hợp, độ tin cậy và yếu tố chắc chắn đều là 1, có nghĩa là khi người dùng truy cập các trang khởi tạo thì chắc chắn sẽ ghé thăm trang được tham chiếu.

Sử dụng các phương pháp khai phá dữ liệu trong các lĩnh vực khác nhau như luật kết hợp, phân tích, thống kê, phân tích địa chỉ trang khởi tạo, phân lớp và phân cụm để khai phá ra các mẫu của người dùng.

Hầu hết địa chỉ của các trang khởi tạo được bố trí theo đồ thị vật lý của trang Web. Mỗi nút là một trang, mỗi cạnh là một đường liên kết giữa các trang. Thông qua việc phân tích đường dẫn trong quá trình truy cập của người dùng có thể tìm ra được mối quan hệ trong việc truy cập của người dùng giữa các đường dẫn (trang web) liên quan.

Ví dụ: Một công ty có địa chỉ Web http://company.com, và các liên kết của nó:

http://company.com/new;

http://company.com/product2;

http://company.com/product1;

http://company.com/products.

Quá trình phân tích logfile cho thấy:

- 70% các khách hàng truy cập vào http://company.com/product2 đều xuất phát từ http://company.com/ thông qua http://company.com/new, http://company.com/products và http://company.com/producttt1.

- 80% khách hàng truy cập vào WebSite bắt đầu từ http://company/products.

- 65% khách hàng rời khỏi site sau khi thăm 4 hoặc ít hơn 4 trang.

Quá trình tích phân cụm dữ liệu cho thấy thông thường các khách hàng được nhóm theo các phần tử dữ liệu giống nhau hoặc có các đặc tính tương tự như nhau. Khi đó, nó trợ giúp cho việc phát triển và thực hiện các chiến lược tiếp thị khách hàng cả về trực tuyến và không trực tuyến cũng như việc trợ giúp trả lời tự động cho khách hàng thuộc cùng nhóm chắc chắn. Khi đó, hệ thống sẽ tạo ra sự thay đổi linh động hơn đối với mỗi Website riêng biệt cho từng khách hàng cụ thể.

4. KẾT LUẬN

Trong bài báo, chúng tôi đã xem xét các khía cạnh chính của khai phá Website tập trung vào khai phá sử dụng Website. Chúng tôi cũng chỉ ra ứng dụng logic mờ để phân tích thông tin của các tập tin nhật ký Webserver sử dụng luật kết hợp mờ.

Một khía cạnh quan trọng khác trong bài báo là sự cá nhân hóa, trong đó các hành vi sử dụng được mô hình hóa bởi hồ sơ, trong đó hầu hết các phần tử này không chính xác. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu phát hiện tấn công website thông qua phân tích tập tin nhật ký sử dụng kỹ thuật khai phá phân cụm mờ kết hợp các luật mờ.

TÀI LIỆU THAM KHẢO

[1]. Agrawal, R., Imielinski, T., Swami, A.: *Mining association rules between sets of items in large databases*. In: Proceedings of the 1993, ACM SIGMOD Conference, pp.207–216 (1993)

[2]. Arotaritei, D., Mitra, S.: Web Mining: a survey in the fuzzy framework. Fuzzy Sets and Systems (2000)

[3]. Au, W.H., Chan, K.C.C.: *An effective algorithm for discovering fuzzy rules in relational databases*. In: Proc. Of IEEE International Conference on Fuzzy Systems, vol. II, pp. 1314–1319 (1998)

[4]. Carbonell, J., Carven, M., Fienberg, S., Mitchell, T., Yang, Y.:*Report on the conald workshop on learning from text and the web.* In: CONALDWorkshop on Learning from Text and The Web (June 1998)

[5]. Cernuzzi, L., Molas, M.L.: *Integrando diferentes Técnicas de Data Mining en procesos de Web Usage Mining* (2003)

[6]. Cooley, R., Mobasher, B., Srivastava, J.: *Web mining: Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns,* pp. 1–11 (2000)

[7]. Chakrabati, S.: *Data Mining for hypertext: A tutorial survey.* ACM SIGKDD Explorations 1(2), 1–11 (2000)

[8]. Delgado, M., Sánchez, D., Vila, M.A.: *Fuzzy cardinality based evaluation of quantified sentences.* Int. J. Aprox.Reasoning 3, 23 (2000)

[9]. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A.: *Mining Text Data: Special Features and Patterns. Pattern Detection and Discovery.* In: Hand, D.J., Adams, N., Bolton, R. (eds.) Proceedings ESF Exploratory

Workshop. Lecture Notes in Artificial Intelligence Series, pp. 140–153 (2002)

[10]. Delgado, M., Marín, N., Sánchez, D., Vila, M.A.: *Fuzzy Association Rules: General Model and Applications.* IEEE Transactions on Fuzzy Systems 11, 214–225 (2003)

[11]. ECML/PKDD Conference 2005, Web Site. Porto, Portugal (2005) http://ecmlpkdd05.liacc.up.pt/

[12]. Etzioni, O.: *The World Wide Web: Quagmire or gold mine.* Comunications of the ACM 39, 65–68 (1996)

[13]. Garofalakis, M.N., Rastogi, R., Seshadri, S., Shim, K.: *Data Mining and the web: Past, present nad future.* In: WorkShop on Web information and data managament, pp.43–47 (1999)

[14]. Hong, T.P., Kuo, C.S., Chi, S.C.: *Mining association rules from quantitative data. Intelligent Data Analysis* 3, 363–376 (1999)

[15]. Hūllermeier, E.: *Fuzzy methods in machine learning and data mining: Status and prospects.* Fuzzy Sets and Systems 156(3), 387–406 (2005)

[16]. Bing Liu, Web mining, Springer, 2007.

[17]. Kraft, D.H., Martín-Bautista, M.J., Chen, J., Vila, M.A: *Rules and fuzzy rules in text: concept, extraction and usage.* International Journal of Approximate Reasoning 34, 145–161 (2003)

[18]. Kuok, C.-M., Fu, A., Wong, M.H.: *Mining fuzzy association rules in databases.* SIGMOD Record 27(1), 41–46 (1998)

[19]. Lee, J.H., Kwang, H.L.: *An extension of association rules using fuzzy sets.* In: Proc. of IFSA'97, Prague, Czech Republic (1997)

[20]. Martín-Bautista, M.J., Kraft, D.H., Vila, M.A., Chen, J., Cruz, J.: *User profiles and fuzzy logic for Web retrieval issues.* Soft Computing Journal 6(5), 365–372 (2004)

[21]. Mitra, S., Pal, S.K:*Data Mining in Soft Computing Framework: A Survey.* IEEE Transactions on Neural Networks, 3–14 (2002)

[22]. Mobasher, B.: Web Usage Mining and Personalization. In: Singh, M.P. (ed.) Practical Handbook of Internet Computing, CRC Press, Boca Raton (2005)

[23]. Nasraoui, O., Frigui, H., Joshi, A., Krishnappuram, R.: *Mining Web accses logs using relational competitive fuzzy clustering.* In: Proceedings of springs Symposium On Natural Language Proccesing Form the www, Stanford, California. March 1997 (1997)

[24]. Wong, C.: Shiu, S. and Pal, S.: *Mining Fuzzy Association Rules for Web Access Case Adaptation.* In: Workshop Proceedings of Soft Computing in Case-Based Reasoning Workshop, in conjunction with the 4th International Conference in Case-Based Reasoning, Vancouver, Canada, pp. 220 (2001)

[25]. Zadeh, L.: *The concept of linguistic variable and its application to approximate reasoning* In Information Sciences 8, 199–251 (1975)

[26]. Justicia et al., 2004. Justicia, C., Martín-Bautista, M. J., Sánchez, D.: Minería de textos: Aplicaciones con lógica difusa. Actas del Congreso Espãnol de Tecnologías con Lógica Difusa, Jaén (In Spanish) (2004).

ABSTRACT

LOGFILE ANALYSIS USING FUZZY LOGIC AND MINING TECHNIQUE

In recent years, with the amounts of website data increasing rapidly on the Internet, the field of website data mining research is very interested. In this paper, we investigate some technical overview of the fuzzy logic and mining techniques used to exploit the website based on analysis of log files – record of user activity while interacting with the Website. In the research, fuzzy clustering techniques and combinations fuzzy clustering rule to improve the efficiency of verifying log dataset from webserver are also performed.

Keywords: Web mining, Fuzzy Logic, File log, Fuzzy.

Nhận bài ngày 06 tháng 12 năm 2016
Hoàn thiện ngày 19 tháng 01 năm 2017
Chấp nhận đăng ngày 01 tháng 5 năm 2017

Địa chỉ: [1] Học viện Kỹ thuật quân sự ;

[2] Phòng Thí nghiệm trọng điểm ATTT- Cục CNTT.

[*] Email: nguyenvanquan87@mail.ru