HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Văn Tiến

PHÁT TRIỂN GIẢI PHÁP THU THẬP VÀ PHÂN TÍCH LOG TRUY CẬP WEBSITE SỬ DỤNG HỌC KHÔNG GIÁM SÁT

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **GS. TS. Từ Minh Phương**

Phản biện 1: PGS. TS. Trần Đăng Hưng

Phản biện 2: TS. Ngô Xuân Bách

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 8 giờ 20 phút ngày 11 tháng 01 năm 2020

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Hiện nay, số lượng website trên toàn cầu là rất lớn, lên tới 1,24 tỉ website (tính đến năm 2018), và số lượng website phát triển thêm hàng nghìn mỗi ngày. Dữ liệu truy cập các trang web với số lượng người dùng khổng lồ chứa rất nhiều thông tin. Các máy chủ lưu trữ website đã có giải pháp ghi log truy cập website. Log truy cập website là một bảng ghi nhật ký truy cập từ tất cả người dùng tương tác với website. Thông thường, việc ghi nhật ký website tại phía máy chủ nhằm mục đích phân tích, đánh giá lưu lượng truy cập website để kiểm soát hiệu năng của hệ thống, chống xâm nhập bất thường phục vụ bảo mật máy chủ web.

Cụ thể, luận văn tập trung vào hai vấn đề chính: 1) nghiên cứu phát triển giải pháp ghi lại tương tác của người dùng với nội dung trên website như mở trang, click vào đường link trên trang, click vào nút trên trang web v.v. ; 2) xác định các nhóm người dùng có nhu cầu thông tin tương tự nhau dựa trên log tương tác ghi lại ở nội dung 1. Thông tin về nhóm người dùng được hiển thị trực quan và có thể sử dụng để phân tích về đối tượng sử dụng website, từ đó cải thiện cấu trúc và nội dung website. Hai vấn đề nghiên cứu trong luận văn là hai bài toán riêng của phân tích dữ liệu Web (Web data mining) nói chung.

Việc đưa ra một giải pháp thu thập và phân tích log website từ phía người dùng là một vấn đề vô cùng quan trọng. Một trong những kỹ thuật được sử dụng phổ biến hiện nay và mang lại hiệu quả cao là kỹ thuật học không giám sát. Đề tài luận văn này sẽ tập trung vào tìm hiểu kỹ thuật tư vấn này, dựa trên hành vi duyệt website của người dùng nhằm đưa ra các phân tích để tư vấn cho người quản trị website có thể nắm bắt được nhu cầu, xu hướng của người dùng website của mình. Từ đó người quản trị sẽ thực hiện các thay đổi website trở nên khoa học hơn, thú vị hơn với người dùng.

Luận văn bao gồm ba chương chính với nội dung như sau:

- Chương 1: Tìm hiểu bài toán thu thập và phân tích log truy cập, giới thiệu tổng quan về khai phá dữ liệu, tổng quan về các giải pháp thu thập, phân tích log truy cập website.

- Chương 2: Trình bày phương pháp thu thập log và phương pháp phân tích log truy cập website sử dụng kỹ thuật phân cụm dữ liệu.

- Chương 3: Thực nghiệm và kết quả: Thử nghiệm triển khai phương pháp thu thập log và cài đặt thuật toán dựa trên kỹ thuật học không giám sát trên bộ dữ liệu thu thập được.

CHƯƠNG 1 - TỔNG QUAN VỀ LOG TRUY CẬP WEBSITE

1.1. Bài toán thu thập và phân tích log truy cập website

Log truy cập hay nhật ký, hoặc vết truy cập (gọi tắt là log) là một danh sách các bản ghi mà một hệ thống ghi lại khi xuất hiện các yêu cầu truy cập các tài nguyên của hệ thống.

Log truy cập website (gọi tắt là web log) chứa tất cả các yêu cầu truy cập các tài nguyên của một website. Các tài nguyên của một website như các file ảnh, các mẫu định dạng và file mã Javascript. Khi một người dùng ghé thăm một trang web để tìm một sản phẩm, máy chủ web sẽ tải xuống thông tin và ảnh của sản phẩm và log truy cập sẽ ghi lại các yêu cầu của người dùng đến các tài nguyên thông tin và ảnh của sản phẩm.

Thu thập log truy cập website là quá trình ghi lại các tương tác của người dùng với website, ví dụ như:

- Xem trang web
- Click vào đường dẫn, nút trên trang web
- Cuộn chuột trên trang web
- Điền dữ liệu vào biểu mẫu, tìm kiếm, …

Bài toán phân tích log truy cập website là một bài toán thuộc lĩnh vực khai phá dữ liệu có:

- Đầu vào: Các bản ghi dữ liệu truy cập hệ thống về hành vi người dùng.

- Đầu ra: Các kết quả phân tích về hệ thống làm cơ sở để đánh giá, cải thiện chất lượng của website.

Để giải quyết hai bài toán trên, chúng ta cần phải tìm hiểu các phương pháp thu thập và phân tích log hiện nay, xem xét các ưu, nhược điểm các phương pháp để lựa chọn các phương pháp phù hợp.

1.2. Các phương pháp thu thập log.

1.2.1. Phương pháp thu thập log phía máy chủ

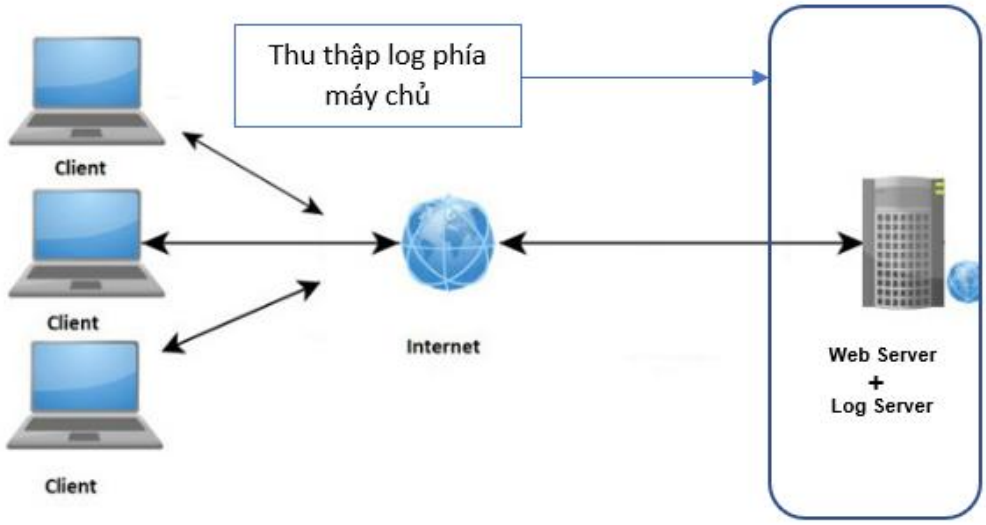
Các phần mềm Web server cho phép lưu lại lịch sử tương tác (log tương tác) giữa người dùng với website. Cụ thể khi trình duyệt gửi yêu cầu của người dùng về máy chủ, các thao tác này được ghi lại trong file log. Hình 1.1 là ví dụ một đoạn log như vậy.

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Hình 1.1: Dữ liệu log thu thập trên máy chủ

Log phía máy chủ web là một nguồn quan trọng để thực hiện khai thác sử dụng web bởi vì từng bản ghi log sẽ được lưu trữ lại cùng những thông tin về người dùng web được cung cấp bởi trình duyệt. Dữ liệu được ghi trong nhật log máy chủ phản ánh việc truy cập (có thể đồng thời) của trang web bởi nhiều người dùng khác nhau. Những tập tin log có thể được lưu trữ dưới định dạng chung hoặc dạng mở rộng.

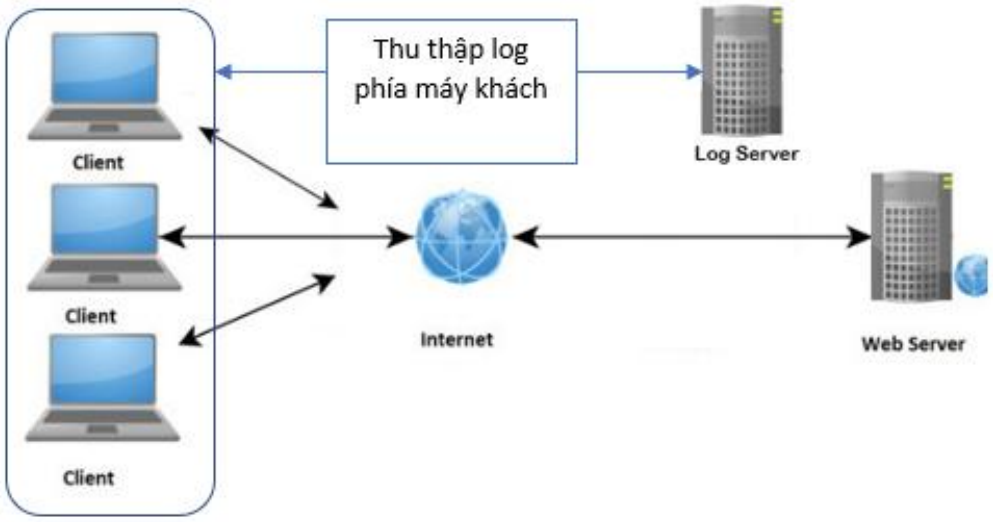
Ưu điểm của phương pháp thu thập log phía máy chủ là thường đi kèm các bộ cài đặt máy chủ web, người quản trị không cần cài đặt thêm phần mềm bên thứ ba, cũng không cần thay đổi mã nguồn website cả phía backend và frontend. Tuy nhiên, cũng có nhiều công cụ được phát triển sẵn với nhiều tính năng nâng cao cho việc thu thập log truy cập.



Hình 1.2: Mô hình thu thập log phía máy chủ

1.2.2. Phương pháp thu thập log phía máy khách

Thu thập log ở phía máy có thể được cài đặt và bằng cách sử dụng các mã hỗ trợ bởi trình duyệt (như Javascripts hoặc Java applets) hoặc bằng cách thay đổi mã nguồn có sẵn của trình duyệt(như Mosaic hay Mozilla) để tăng cường khả năng thu thập dữ liệu. Việc cài đặt thu thập dữ liệu log phía máy khách đòi hỏi phải có sự hợp tác từ phía người dùng, họ cần phải bật chức năng cho phép JavaScripts hay Java applets.



Hình 1.3: Mô hình thu thập log phía máy khách

Trong luận văn này sẽ giới thiệu 2 phần mềm thu thập log phía máy khách là Google Analytics (do Google phát triển) và Countly (Mã nguồn mở - có thể tự cài đặt)

1.2.2.1. Phần mềm thu thập log Google Analytics

Google Analytics là một dịch vụ phân tích trang web miễn phí cung cấp cho người quản trị các công cụ để đo lường sự thành công của trang web liên quan đến tiếp thị, tối ưu hóa nội dung hoặc thương mại điện tử.

Google Analytics sử dụng kết hợp các cookie và phiên tạm thời để theo dõi hành vi trực tuyến của khách truy cập. Google Analytics sử dụng cookie của bên thứ nhất để xác định duy nhất từng khách truy cập. Bằng cách truy cập trang web, khách truy cập kích hoạt JavaScript này, thông tin cookie sẽ được chuyển đến tài khoản Google Analytics của người quản trị.

1.2.2.2. Phần mềm thu thập log Countly

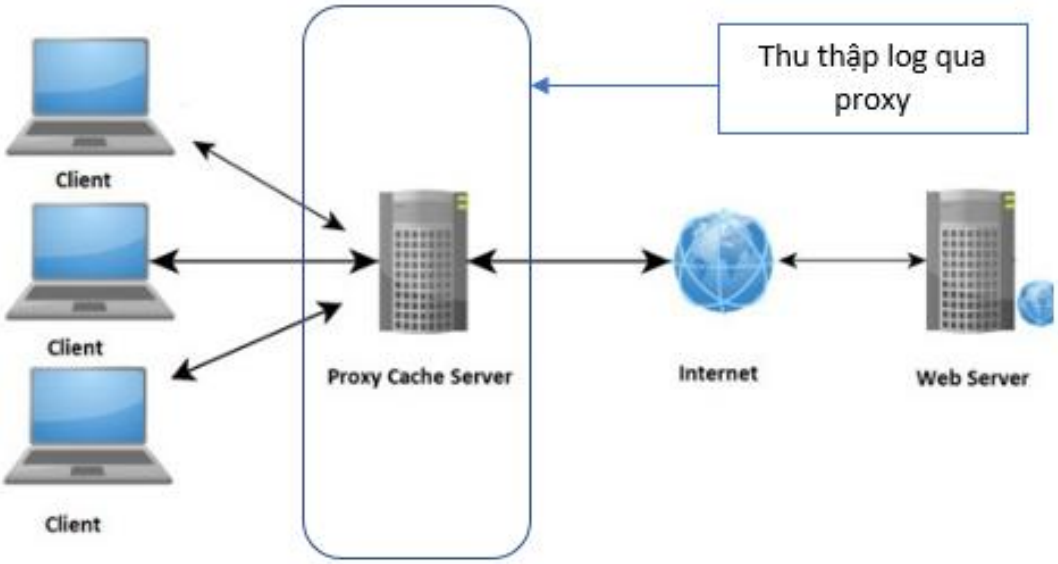
Countly là phần mềm phân tích web, ứng dụng nguồn mở được viết bằng NodeJS và sử dụng cơ sở dữ liệu MongoDB, Countly có thể so sánh với Google Analytics, mặc dù Countly là phần mềm máy chủ mà bất kỳ ai cũng có thể cài đặt và chạy trên máy chủ của riêng họ, trong khi Google Analytics là dịch vụ phần mềm do Google cung cấp. Ứng dụng này giúp quản trị viên theo dõi và quan sát luồng lượt xem trên trang web. Là một khung phân tích web chung, Counlty có thể được mở rộng để theo dõi và phân tích bất kỳ ứng dụng web nào.

Cả Google Analytics và Countly đều có những điểm mạnh riêng, đều hỗ trợ khả năng ghi lại tương tác của người dùng trực tiếp trên website. Google Analytics và Countly đều có các báo cáo về lưu lượng, các phân tích về hành vi người dùng, báo cáo theo thời gian thực với rất nhiều thông tin thu thập được từ người dùng. Tuy nhiên, với Google Analytics, cáo báo cáo, các thuật toán là do Google phát triển và thêm vào các tính năng theo thời gian, còn đối với Countly, do là mã nguồn mở, nên có độ tùy biến cao hơn. Chúng ta hoàn toàn có thể chủ động phát triển thêm các tính năng để thêm vào hệ thống đang có.

1.2.3. Phương pháp thu thập log qua proxy

Máy chủ proxy hoạt động như một cổng nối giữa người dùng và Internet. Đây là một máy chủ trung gian giữa người dùng cuối và trang web họ truy cập. Các máy chủ proxy cung cấp các chức năng, bảo mật và riêng tư khác nhau phụ thuộc vào nhu cầu của quản trị viên hoặc chính sách công ty.

Thu thập log thông qua proxy được thực hiện ở máy chủ trung gian. Phương pháp này có thể thu thập được các yêu cầu duyệt web từ phía máy khách. Tuy nhiên, các hành vi của người dùng như nhấp chuột, hay cuộn chuột thì vẫn không thu thập được. Hiệu suất của proxy phụ thuộc nhiều vào khả năng dự đoán chính xác các yêu cầu duyệt web của người dùng trong tương lai. Phân tích log truy cập qua proxy chủ yếu nhằm giúp cải thiện hiệu suất của proxy để giảm giá thành chi phí Internet trong nội bộ của công ty, tổ chức. Hình 1.4 cho thấy cách hoạt động của phương pháp thu thập log thông qua proxy.



Hình 1.4: Mô hình thu thập log qua proxy

Trong các giải pháp trên, để thực hiện khai phá dữ liệu hành vi người dùng trang web thì giải pháp thu thập log phía máy khách là phù hợp nhất với nhiều tiêu chí như dữ liệu có tính thực tế cao, chi phí triển khai thấp hơn so với các giải pháp còn lại.

1.3. Phương pháp phân tích log

Có nhiều phương pháp phân tích log truy cập khác nhau, tùy vào mục đích phân tích có độ phức tạp khác nhau. Ví dụ chỉ cần đưa ra các thống kê về lượt xem, giờ xem thì có thể sử dụng các phương pháp thống kê đơn giản rồi sử dụng các dạng bảng biểu, biểu đồ để thể hiện. *Luận văn sẽ tập trung vào việc xác định các nhóm người dùng có nhu cầu thông tin tương tự nhau*. Việc xác định nhóm người dùng được thực hiện bằng phương pháp phân cụm - một phương pháp học máy không giám sát.

1.3.1. Giới thiệu học không giám sát

Học không giám sát (Unsupervised Learning) là một nhóm thuật toán học máy được phân chia bằng phương thức học. Trong thuật toán này, chúng ta không biết được kết quả đầu ra hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán học không giám sát sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân cụm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

Các bài toán học không giám sát được chia thành hai loại:

- Phân cụm (clustering)

- Học luật kết hợp (association rule mining

1.3.2. Một số kỹ thuật phân cụm dữ liệu

Ta có thể khái quát hóa khái niệm Phân cụm dữ liệu: Phân cụm dữ liệu là một kỹ thuật trong khai phá dữ liệu, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên, tiềm ẩn, quan trọng trong tập dữ liệu lớn từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định.

Như vậy, phân cụm dữ liệu là quá trình phân chia dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong cụm tương tự nhau với nhau và các phần tử trong các cụm khác nhau sẽ không tương tự với nhau. Số các cụm dữ liệu được phân có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định của phương pháp phân cụm.

Trong học máy, Phân cụm dữ liệu được coi là thuật toán học không giám sát, một số kỹ thuật phân cụm phổ biến thường được sử dụng là: phân cụm phân hoạch, phân cụm phân cấp và phân cụm theo mật độ.

1.3.2.1. Phân cụm phân hoạch

Giới thiệu về phân cụm phân hoạch

1.3.2.2. Phân cụm theo mật độ

Giới thiệu về phân cụm theo mật độ

1.3.2.3. Phân cụm phân cấp

Giới thiệu về phân cụm phân cấp, áp dụng phân cụm phân cấp trong phân tích log.

1.4. Kết luận chương

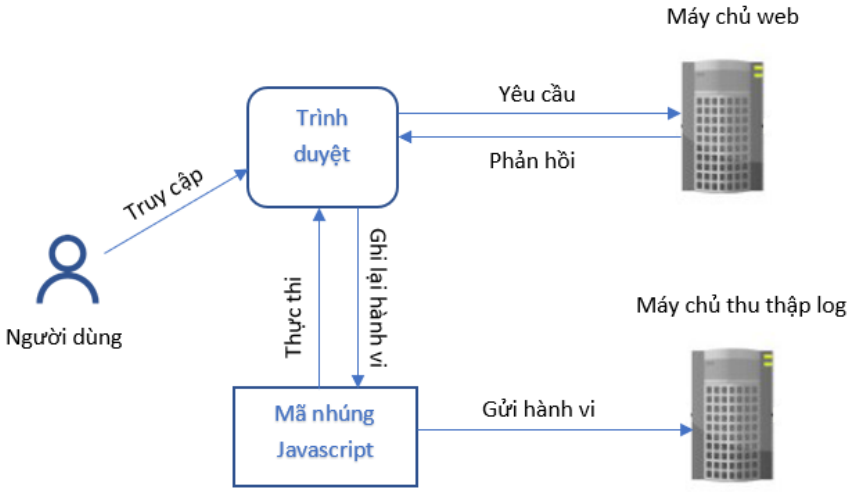
Chương 1 đã trình bày về khái niệm log truy cập, bài toán thu thập và phân tích log truy cập. Chương cũng giới thiệu về tổng quan về các giải pháp thu thập log và kỹ thuật phân tích log bằng phương pháp học không giám sát.

CHƯƠNG 2 - CÁC KỸ THUẬT PHÂN TÍCH LOG TRUY NHẬP

2.1. Xây dựng công cụ thu thập log

Ngày nay, các công cụ phân tích website được cải tiến không ngừng. Nó hỗ trợ cho người quản trị website có thể nắm được các số liệu thống kê, phân tích về website của mình. Một số công cụ còn dựa vào cookies, thông tin của trình duyệt, kết hợp với kho dữ liệu khổng lồ của họ để xác định độ tuổi, giới tính, sở thích của người dùng để đưa ra các phân tích chuyên sâu nhằm tối ưu về lợi nhuận bán hàng cho các trang thương mại điện tử.

Trong chương 1, ta đã xem xét các đặc điểm của các giải pháp thu thập log. Trong các giải pháp, thu thập log phía máy khách có nhiều ưu điểm phù hợp cho việc thu thập log truy cập phục vụ cho quá trình khai phá dữ liệu phân cụm người dùng.



Hình 2.1: Sơ đồ mô tả hoạt động hệ thống thu thập log

Hình 2.1 mô tải quá trình hoạt động của một hệ thống thu thập log hoàn chỉnh khi người dùng truy cập vào website.

Như vậy, Cần phải cài đặt thêm phần mềm trên máy chủ thu thập log, phần mềm này có khả năng sinh ra mã nhúng Javascript để tích hợp vào máy chủ web hiện có. Qua khảo sát một số phần mềm hỗ trợ thu thập log phía máy khách, Countly là một chương trình mã nguồn mở được xây dựng trên ngôn ngữ NodeJS với nhiều tính năng nổi bật. Tuy nhiên công cụ này được xây dựng để phân tích, thống kê các dữ liệu duyệt web cơ bản của người dùng. Do đó dữ liệu log không được lưu lại mà chỉ phục vụ cho việc tính toán, thống kê theo từng giai

đoạn. Để có thể thu thập một số lượng bản ghi đủ dùng cho thuật toán khai phá dữ liệu, cần phải phát triển thêm mã nguồn của Countly.

Ban đầu, Countly chỉ lưu lại 1000 bản ghi log truy cập website gần nhất cho mỗi website được theo dõi trên Countly. Do giới hạn lưu trữ, không thể lưu toàn bộ dữ liệu log truy cập, đối với các trang web có số lượng truy cập lớn số lượng bản ghi có thể tăng rất nhanh dẫn đến việc quá tải và làm Countly ngừng hoạt động. Số lượng bản ghi lưu lại cần được tính toán, cân đối phù hợp với cấu hình của máy chủ hoặc thiết lập sao lưu sang máy chủ khác để đảm bảo hoạt động của máy chủ thu thập dữ liệu.

2.2. Xây dựng đồ thị tương tự

Sau khi thu thập log, trên dữ liệu thống kê có danh sách các người dùng đã truy cập website. Tập hợp người dùng này được coi là một nhóm người dùng lớn. Mỗi người dùng đều có các mối quan tâm, sở thích khác nhau. Tuy nhiên sẽ có nhiều người dùng lại có sở thích, mối quan tâm tương đồng nhau. Việc đánh giá sở thích, mối quan tâm của người dùng trên một tập hợp người dùng có nhiều điểm khác nhau là rất khó khăn. Muốn tìm hiểu được mối quan tâm của người dùng với website, ta phải chia nhóm người dùng lớn này thành các nhóm người dùng nhỏ hơn, mỗi thành viên của một nhóm người dùng sẽ có các sở thích tương tự với nhau trong cùng nhóm, và mỗi nhóm khác nhau sẽ có các mối quan tâm khác nhau.

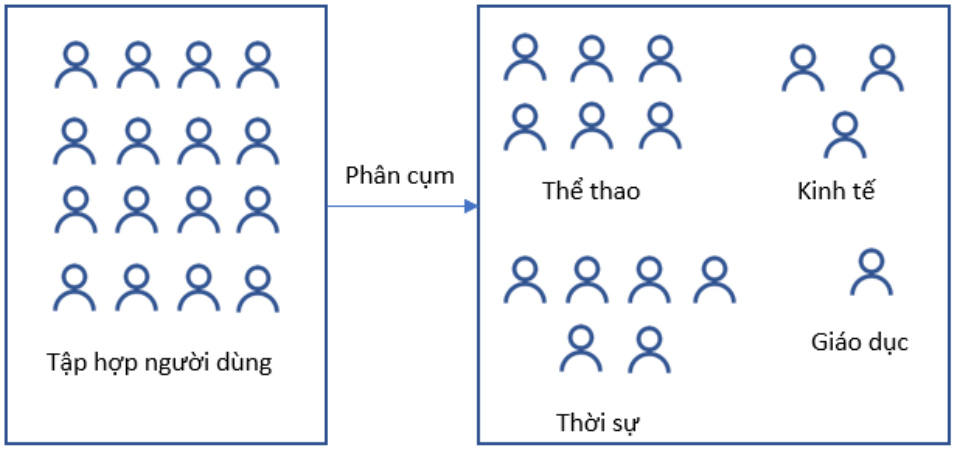
Trong phạm vi luận văn, hai người dùng được coi là có sở thích giống nhau nếu cùng xem các thông tin giống nhau. Thông tin được xác định ở các mức khác nhau. Cụ thể, hai người dùng được coi là tương tự nếu:

a. Cùng xem những trang Web giống nhau

b. Cùng xem những trang Web thuộc thể loại giống nhau

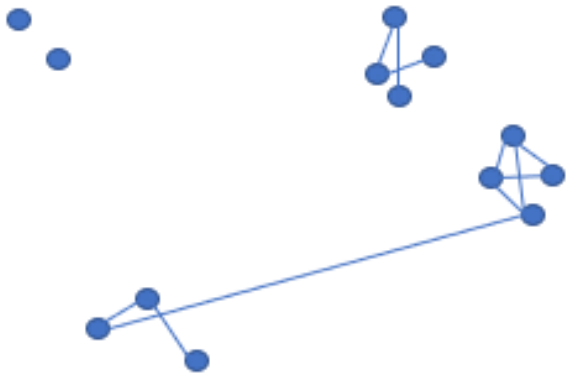
c. Cùng xem những trang Web về các chủ để giống nhau

Sau khi xác định được độ tương tự giữa từng đôi người dùng, có thể sử dụng kỹ thuật phân cụm để xác định các nhóm người dùng cùng sở thích. Phân cụm dữ liệu là một phương pháp học máy không giám sát đã được giới thiệu ở chương 2. Hình 2.2 minh họa cho quá trình phân cụm người dùng.



Hình 2.2: Hình minh họa phân cụm người dùng

Dữ liệu log thu thập được lưu trữ dưới dạng các bản ghi, mỗi bản ghi thể hiện thao tác ghé thăm một trang web của người dùng hoặc hành vi của người dùng trên trang web như cuộn trang web, click vào các đường dẫn, hình ảnh, ... Phân cụm người dùng là quá trình xác định các nhóm người dùng có điểm giống nhau, vì vậy cần biểu diễn dữ liệu dưới dạng đồ thị thể hiện sự tương tự giữa người dùng trong hệ thống (gọi tắt là đồ thị tương tự). Do đó cần phải xử lý dữ liệu bản ghi tuần tự này để chuyển dữ liệu sang dạng đồ thị. Hình 2.4 cho thấy ví dụ về một đồ thị đơn giản thể hiện mối tương tự của người dùng. Đỉnh của đồ thị đại diện cho người dùng, cạnh giữa hai đỉnh thể hiện độ tương tự giữa hai người dùng.



Hình 2.3: Đồ thị vô hướng thể hiện độ tương tự của người dùng

Quá trình này xây dựng đồ thị tương tự gồm các bước: Loại bỏ các bản ghi dư thừa, Xác định chủ đề cho các trang web, Xác định độ tương tự của người dùng.

2.2.1. Loại bỏ các bản ghi dư thừa

Dựa vào đặc điểm nội dung trang web, các dữ liệu cần thiết để loại bỏ các dữ liệu cần thiết giúp tiết kiệm thời gian xử lý.

2.2.2. Xác định các chuyên mục, chủ đề

Để phân tích, đánh giá được kết quả phân cụm, cần xác định được chuyên mục, thể loại của các trang web. Ví dụ, nhóm các trang web về tin tức thể thao, chính trị, … Một số website có hệ thống chuyên mục được xác định sẵn, một số website không phân các trang web vào các chuyên mục cố định trước.

Đối với những trang web không được chia các chuyên mục cố định, ta có thể sử dụng thuật toán LDA (Latent Dirichlet Allocation) để xác định các chủ đề cho mỗi trang web.

2.2.3. Xác định độ tương tự của người dùng

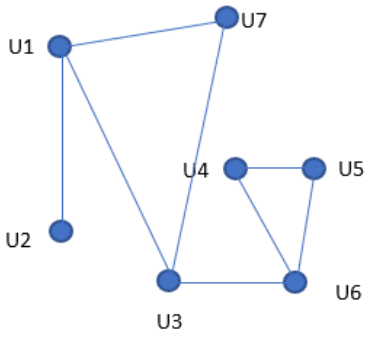
Có thể có nhiều cách xác định độ tương tự giữa hai người dùng. Ví dụ: Có thể dựa vào chuỗi các sự kiện tương tác của người dùng, hoặc dựa vào số lượt ghé thăm cùng một trang web giữa hai người dùng. Trong luận văn này sử dụng số lượt ghé thăm cùng một trang web để làm cơ sở xác định độ tương tự giữa hai người dùng.

Sau khi đã chuẩn hóa dữ liệu các bản ghi và chuẩn bị các dữ liệu cần thiết, ta biểu diễn dữ liệu này dưới dạng đồ thị tương tự.

Đồ thị theo trang web: Trọng số của đồ thị là giá trị $sim_{page\_visit}(u_p, u_q)$.

Đồ thị theo trang chuyên mục: Trọng số của đồ thị là giá trị $sim_{cate\_visit}(u_p, u_q)$.

Đồ thị theo trang chủ đề: Trọng số của đồ thị là giá trị $sim_{topic\_visit}(u_p, u_q)$.



Hình 2.4: Ví dụ về đồ thị tương tự của người dùng

2.3. Phân cụm người dùng

Sau khi xác định được độ tương tự của người dùng, xây dựng được đồ thị mối quan tâm tương đồng của người dùng, bước tiếp theo là phân cụm người dùng có cùng mối quan tâm bằng cách phân cụm các đồ thị tương tự.

Phương pháp phân cụm này xác định mối quan tâm của người dùng từ dữ liệu hành vi duyệt web của người dùng. Phương pháp này sẽ phân cụm người dùng thành các nhóm người dùng có cùng mối quan tâm, sở thích bằng cách phân cụm đồ thị tương tự, trong đó các đỉnh là người dùng và các cạnh thể hiện sự tương đồng trong hành vi duyệt web giữa hai người dùng đã được xây dựng ở phần trên.

2.4. Xác định ý nghĩa các cụm người dùng

Sở thích của người dùng có thể được suy ra ý nghĩa từ các cụm trong từng loại đồ thị tương tự.

Đồ thị theo chuyên mục: Đối với mỗi cụm, số lượng người dùng truy cập từng chuyên mục được tính toán. Sau đó, chọn các chuyên mục N_c đầu tiên, có số lượng người dùng truy cập lớn nhất cho mỗi cụm. Tên của các chuyên mục N_c này đại diện cho mối quan tâm của người dùng của cụm.

Đồ thị theo chủ đề: Đối với mỗi cụm, số lượng người dùng truy cập từng chủ đề được tính toán. Theo mô hình LDA, mỗi chủ đề là sự kết hợp của các từ khóa và mỗi từ khóa đóng góp một trọng số nhất định cho chủ đề. Ta chọn K từ khóa cho mỗi chủ đề. Sau đó, chúng ta chọn N_t các chủ đề phổ biến nhất có số lượng người dùng truy cập lớn nhất cho mỗi cụm. Sự kết hợp của các từ khóa từ các chủ đề được chọn N_t này có thể đóng vai trò giải thích cho mối quan tâm của người dùng.

Đồ thị theo trang web: Đối với mỗi cụm trong đồ thị trang, chúng ta xác định nhóm trang mà người dùng đã truy cập và số lượng người dùng truy cập vào nhóm này. Không thể khám phá mối quan tâm của người dùng trong mỗi cụm bằng cách suy luận trực tiếp từ tập hợp các trang. Tuy nhiên, công việc này có thể được thực hiện thông qua tập hợp các chuyên mục hoặc chủ đề của trang. Một trang web có thể được thêm vào một hoặc nhiều chuyên mục. Và bằng cách sử dụng mô hình chủ đề dựa trên LDA, chúng ta có thể phân loại tiêu đề của một trang thành một chủ đề cụ thể. Sau khi xác định bộ chuyên mục hoặc bộ chủ đề và số lượng người dùng được truy cập qua các trang, chúng ta có thể phân tích và hiểu sở thích của người dùng của cụm, tương tự như các trường hợp đồ thị chuyên mục và đồ thị theo chủ đề.

Các chi tiết về phương pháp phân tích mối quan tâm người dùng được đề xuất như thuật toán cụ thể dưới đây.

Algorithm User_Interest_Analysis

Input: *dataset* //Tập dữ liệu đã qua tiền xử lý

Output: None

Procedure User_interest_analysis(*dataset***)**

1. *set_of_titles* ← Trích xuất tập các tiêu đề từ *dataset*;
2. *user_list* ← Trích xuất danh sách người dùng từ *dataset*;
3. *LDA_Model* ← Xây dựng LDA Model từ *set_of_titles*;
4. **for** each user u_i and u_j in *user_list* {
5. *page_graph*(u_i , u_j) = $\text{sim}_{page\_visit}(u_i,u_j)$; // Xây dựng đồ thị tương tự theo trang web
6. *cate_graph*(u_i , u_j) = $\text{sim}_{cate\_visit}(u_i,u_j)$; // Xây dựng đồ thị tương tự theo chuyên mục
7. *topic_graph*(u_i , u_j) = $\text{sim}_{topic\_visit}(u_i,u_j)$; // Xây dựng đồ thị tương tự theo chủ đề được xác định bằng *LDA_Model*
8. } //**end for**
9. *Page_SubClusters*←Clustering(*page_graph*);// Phân cụm đồ thị thành các cụm phân cấp
10. *Cate_SubClusters*←Clustering(*cate_graph*);
11. *Topic_SubClusters*←Clustering(*topic_graph*);
12. $Topic\_Terms_{Page}$, $Cate\_Terms_{Page}$ ← Trích xuất từ khóa từ *Page_SubClusters* bằng cách gán tên chuyên mục và chủ đề cho mỗi trang web sử dụng *LDA_Model*;
13. $Terms_{Cate}$ ← Trích xuất các chuyên mục từ *Cate_SubClusters*;
14. $Terms_{Topic}$ ← Trích xuất các chủ đề từ *Topic_SubClusters*;
15. Hiển thị $Topic\_Terms_{Page}$, $Cate\_Terms_{Page}$, $Terms_{Cate}$, $Terms_{Topic}$ kết quả phan cụm;

End Procedure

2.4. Kết luận chương

Chương 2 đã trình bày về cách xây dựng giải pháp kỹ thuật thu tập log, phân tích log. Chương cũng đã trình bày cụ thể về các bước chi tiết trong phương pháp xác định nhóm người dùng có điểm giống nhau và cách xác định mối quan tâm của người dùng từ các nhóm này.

CHƯƠNG 3 - CÀI ĐẶT VÀ THỬ NGHIỆM

3.1. Cài đặt công cụ thu thập log truy cập website

3.1.1. Yêu cầu hệ thống

Countly được thiết kế để chạy trên máy chủ Linux do đó không hỗ trợ các nền tảng hệ điều hành khác như Microsoft Windows hoặc MacOS. Một số hệ điều hành được hỗ trợ như:

- Ubuntu 16.04, 18.04, 18.10 (không bao gồm Ubuntu 19.4)
- Red Hat Enterprise Linux 6.9 trở lên (không bao gồm RHEL 8.0)
- CentOS Linux 6.9 trở lên

Countly cũng chỉ hỗ trợ hệ điều hành có kiến trúc 64bit, yêu cầu môi trường NodeJS 8.x trở lên và MongoDB 3.6.x trở lên

Về phần cứng, Countly yêu cầu máy chủ có tối thiểu 2 CPUs và ít nhất 2GB RAM để có thể hoạt động. Ổ đĩa cứng yêu cầu tối thiểu 20GB.

Trong luận văn này, cho mục đích thử nghiệm, sử dụng máy chủ có cấu hình như sau:

- Hệ điều hành: Ubuntu 16.04
- Phần cứng: CPU: 2 Core, RAM 2GB, SSD 55GB
- Môi trường được cài đặt đầy đủ theo yêu cầu của Countly

3.1.2. Cài đặt hệ thống

Cài đặt môi trường NodeJS 8.x

Cài đặt MongoDB

Sau khi cài đặt môi trường, tiến hành cài đặt công cụ countly lên máy chủ. Các thông số đều được tự động điều chỉnh phù hợp với cấu hình máy chủ đang cài đặt thông qua chức năng cài đặt được cung cấp bởi Countly.

Bước tiếp theo, để có thể thu thập được dữ liệu, cần phải thêm ứng dụng với các thông tin chi tiết. Ứng dụng này để phân biệt giữa các website được quản lý chung trong hệ thống của Countly.

Cuối cùng, cần sinh mã những javascript, mã nhúng này được nhúng trực tiếp lên website cần tích hợp thu thập dữ liệu.

```
<script type='text/javascript'>
//some default pre init
var Tracker = Tracker || {};
Tracker.q = Tracker.q || [];

//provide tracker initialization parameters
Tracker.app_key = 'c64480610999f5141b8f98f7f3df95d9d8f91fe3';
Tracker.url = 'http://207.148.79.97';

Tracker.q.push(['track_sessions']);
Tracker.q.push(['track_pageview']);
Tracker.q.push(['track_clicks']);
Tracker.q.push(['track_scrolls']);
Tracker.q.push(['track_errors']);
Tracker.q.push(['track_links']);
Tracker.q.push(['track_forms']);
Tracker.q.push(['collect_from_forms']);

//load tracker script asynchronously
(function() {
   var cly = document.createElement('script'); cly.type = 'text/javascript';
   cly.async = true;
   //enter url of script here
   cly.src = 'http://207.148.79.97/sdk/web/tracker.min.js';
   cly.onload = function(){Tracker.init()};
   var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(cly, s);
})();
</script><!-- style | dynamic -->
```

Hình 3.1: Mã nhúng tích hợp dành cho website cần thu thập

3.2. Phân tích log truy cập website

3.2.1. Tập dữ liệu thực nghiệm

Trong phạm vi luận văn này, để thực nghiệm xây dựng hệ thống thu thập log và phân tích log truy cập, dữ liệu log được thu thập từ cổng thông tin Học viện Công nghệ Bưu chính Viễn thông (PTIT). Trong tập dữ liệu này, ta thu thập tất cả các hành vi của người dùng và thu thập thông tin của các trang web như chuyên mục và tiêu đề.

Cổng thông tin Học viện Công nghệ Bưu chính Viễn thông là một website được cấu trúc thành nhiều trang web con, mỗi trang web con thuộc một hoặc nhiều chuyên mục. Có tổng số trên 20 chuyên mục riêng biệt, phổ biến như: Thông báo sinh viên, Tin tức, Đào tạo quốc tế, …

Dữ liệu sử dụng để phân tích trong luận văn được thu thập trong 3 tháng (từ 01/04/2019 – 30/06/2019) với khoảng 150,000 bản ghi log tương tác của người dùng. Các thông tin thu thập được bao gồm chi tiết về các hoạt động của người dùng như xem trang, click, tìm kiếm, nội dung của các trang web (bao gồm tiêu đề và nội dung).

Các tác vụ tiền xử lý bao gồm nhận dạng chuyên mục, ước tính thời gian trong khoảng thời gian người dùng dành cho một trang web và làm sạch dữ liệu. Chuyên mục của một bài đăng trong một trang web dễ dàng được xác định bởi trường ID chuyên mục nhưng đôi khi không có chuyên mục trong trang web. Để cải thiện chất lượng dữ liệu, ta xóa các dữ liệu không liên quan không có chuyên mục hoặc rất hiếm khi người dùng truy cập. Trong khoảng thời gian người dùng dành cho một trang web, ta tính toán dựa trên thời gian của hai yêu cầu

web liên tiếp của cùng một người dùng. Các nghiên cứu đã chỉ ra rằng 55% lượt xem trang trên internet kéo dài dưới 15 giây. Thông thường, nó không quá 180 giây.

Thực nghiệm này cũng bỏ qua các trang có lượt xem trang kéo dài ít hơn hoặc bằng 5 giây vì điều đó cho thấy rằng người dùng không có bất kỳ mối quan tâm nào trên các trang này (T = 5). Sau khi tiền xử lý, số lượng hồ sơ được giảm rất nhiều, so với dữ liệu ban đầu. Kết quả là bộ dữ liệu thử nghiệm chứa 5360 người dùng và 19 chuyên mục. Các mô tả chi tiết của dữ liệu nhấp chuột dòng trước và sau khi tiền xử lý được liệt kê trong bảng 3.1

Bảng 3.1: Tập dữ liệu hành vi duyệt web từ website PTIT Portal

Giá trị	Bộ dữ liệu đã lọc
Số bản ghi	63000
Số lượng người dùng	5360
Số lượng chuyên mục	19
Thời gian duyệt web trung bình	12,7 giây
Số lượng trang web	1017

Để xác định các chủ đề cho các trang web, thực nghiệm này sử dụng công cụ LDA từ gói Gensim (https://pypi.org/project/gensim/) . LDA được áp dụng cho tập hợp các tiêu đề được trích xuất từ tất cả các trang web trong bộ dữ liệu. Hai tham số của LDA được nghiên cứu thử nghiệm sử dụng dữ liệu thực là *number_of_topics* (số lượng chủ đề) và *eta*. Trong thực nghiệm này, eta là 0,01. Nó đủ nhỏ để làm cho các chủ đề được cấu thành từ một vài từ. Để dễ dàng hiểu ý nghĩa của một chủ đề, mỗi chủ đề được thể hiện bằng năm từ có thể xảy ra nhất. Và sử dụng thủ tục tìm kiếm lưới, *number_of_topics* là 50 là giá trị tốt nhất. Các giá trị ngưỡng α_{page}, α_{cate} và α_{topic} cũng được thử nghiệm nghiên cứu bằng cách sử dụng bộ dữ liệu này. Trong thực nghiệm này, lần lượt sử dụng trang web là α_{page} 0,003, α_{cate} 0,1 và α_{topic} 0,03. Bởi vì bộ dữ liệu được thu thập từ một cổng web của trường đại học, nó có thể nhóm người dùng thành các nhóm khác nhau như khách truy cập, sinh viên trong trường đại học, sinh viên bên ngoài trường đại học, giảng viên và nhân viên khác của trường đại học. Sau đó, các nhóm người dùng tên này được sử dụng trong phân tích kết quả thực nghiệm.

Với kỳ vọng có thể xác định được các thông tin có ý nghĩa như sở thích của người dùng, đối tượng người dùng nào quan tâm đến các nội dung nào trên cổng thông tin. Dựa trên

các cách tiếp cận khác nhau để phân tích thông tin của người dùng sử dụng cả dữ liệu được gán nhãn (theo chuyên mục) và dữ liệu chưa được gán nhán (theo chủ đề).

3.2.2. Xác định số cụm dữ liệu

Cần phải xác định số cụm phù hợp với dữ liệu người dùng hiện tại. Không phải số cụm lúc nào cũng cố định mà sẽ được tối ưu để phù hợp theo từng giai đoạn. Ví dụ, dữ liệu thu thập được trong hai tháng hiện tại được chia thành 5 cụm sẽ là tối ưu nhất, nhưng trong 2 tháng tiếp theo, có thể cần được chia thành 7 cụm mới phù hợp. *Chỉ số Dunn* (dunn index) được sử dụng để đánh giá kết quả phân cụm. *Chỉ số Dunn* được tính như sau:

$$D = \frac{min.separation}{max.diameter}$$

Trong đó: *min.separation* là khoảng cách nhỏ nhất giữa các cụm khác nhau.

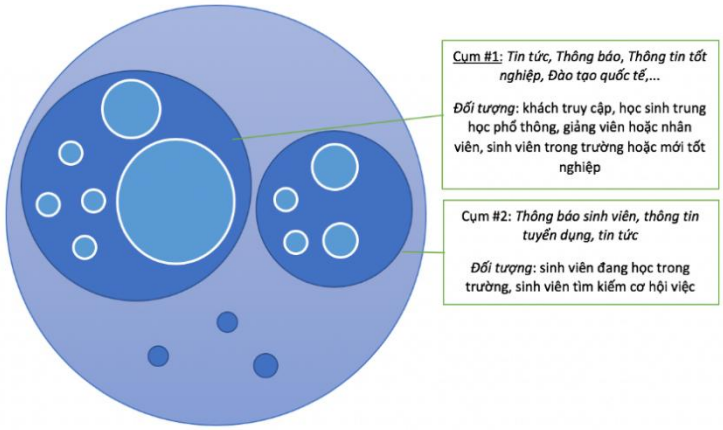
max.diameter là khoảng cách lớn nhất trong nội bộ cụm (giống như đường kính).

Nếu tập dữ liệu chứa các cụm nhỏ gọn và tách biệt, đường kính của các cụm được dự kiến là nhỏ và khoảng cách giữa các cụm được dự kiến sẽ lớn. Do đó, *chỉ số Dunn* nên được tối ưu hóa, giá trị *D* càng lớn thì kết quả phân cụm càng tối ưu.

3.2.3. Kết quả thực nghiệm.

Đối với đồ thị theo chuyên mục, Do một số trang web không được chia vào chuyên mục nào hoặc có những chuyên mục tập trung quá nhiều trang web được loại bỏ, chỉ còn 1857 người dùng trong cụm ban đầu. Sau khi thử nghiệm số chia số cụm ban đầu từ 3 đến 10 cụm, *chỉ số Dunn* tối ưu nhất khi chia thành 5 cụm.

1857 người dùng ban đầu trong bộ dữ liệu được phân thành 5 cụm riêng biệt. Trong số các cụm này, hai cụm hàng đầu về kích thước chứa hơn 600 thành viên. 3 cụm khác bị bỏ qua vì quá nhỏ. Cụm đầu tiên có 3 cụm phụ quan trọng khác nhau và cụm thứ hai chỉ có 2 cụm phụ quan trọng. Dựa trên các kết quả phân cụm theo phân cấp được hiển thị trong Bảng 3.2 và Bảng 3.3, có thể dễ dàng chia người dùng thành 2 nhóm sở thích.



Hình 3.2: Kết quả phân loại người dùng theo chuyên mục

Nhóm đầu tiên quan tâm đến Tin tức từ trường đại học, thông tin tốt nghiệp và đào tạo quốc tế. Nhóm thứ hai quan tâm đến Thông báo sinh viên, Việc làm nhưng không quan tâm đến Tin tức từ trường đại học. Có thể phán đoán rằng người dùng trong nhóm đầu tiên có thể là khách truy cập, giảng viên hoặc nhân viên khác trong trường đại học muốn xem tin tức. Một số là học sinh trung học muốn xem thông tin nhập học và phần còn lại là sinh viên trong trường đại học đã tốt nghiệp hoặc sinh viên xuất sắc đang tìm kiếm đào tạo quốc tế. Người dùng trong nhóm thứ hai có thể là sinh viên bình thường đang học đại học. Những sinh viên này không quan tâm đến tin tức chung từ trường đại học mà chỉ quan tâm đến thông tin liên quan đến sinh viên. Phần còn lại là những sinh viên muốn tìm việc thực tập hoặc công việc. Hình 3.5 cho thấy kết quả phân loại người dùng.

Bảng 3.2: Kết quả phân cụm cấp 1 đồ thị theo chuyên mục

Cụm cấp 1	**Số người dùng**	**Các chuyên mục**
Cluster 1	1250	Tin tức; Thông báo; Thông tin tốt nghiệp, Thông báo văn bằng; Việc làm cho giảng viên; Trao đổi sinh viên
Cluster 2	622	Thông báo cho sinh viên; Thông tin tuyển dụng; Tin tức

Bảng 3.3: Kết quả phân cụm cấp 2 đồ thị theo chuyên mục

Cụm cấp 2	Cụm cha	Số người dùng	Các chuyên mục
Cluster 1	Sub cluster 1	810	Tin tức
	Sub cluster 2	145	Thông tin tốt nghiệp; Thông báo văn bằng; Việc làm cho giảng viên
	Sub cluster 3	127	Thông báo; Tin tức
	Sub cluster 4	33	Trao đổi sinh viên; Đào tạo quốc tế
Cluster 2	Sub cluster 5	527	Thông báo cho sinh viên; Tin tức
	Sub cluster 6	75	Thông tin tuyển dụng; Thông báo sinh viên; Cơ hội việc làm

Phân tích cho thấy một bộ phận người dùng không quan tâm đến tin tức chung chung mà chỉ quan tâm đến tin tức liên quan đến nhiệm vụ học tập và thi cử. Một lý do có thể là không có nhiều tin tức. Trong vòng một tháng, số lượng bài viết mới truy cập là khoảng 1.000. Đây là một thông tin có giá trị cho các quản trị viên cổng thông tin web và các nhà lãnh đạo trường đại học để giúp cải thiện trang web bằng cách cung cấp nhiều thông tin hữu ích hơn.

Đồ thị theo chủ đề, Áp dụng thuật toán phân cụm vào đồ thị chủ đề bằng dữ liệu tiêu đề và nội dung của các trang, người dùng được phân thành 8 cụm. Do kết quả tương tự cho cả hai đồ thị chủ đề, chỉ có kết quả trên đồ thị theo chủ đề dựa trên tiêu đề được trình bày ở đây.

Bảng 3.4: Kết quả phân cụm cấp 1 đồ thị theo chủ đề

Cụm cấp 1	Số người dùng	Chủ đề
Cluster 1	1415	(Thông báo, kết quả, việc làm, điểm chuẩn, chất lượng)
Cluster 2	1097	(Công nghệ, chính quy, bằng tốt nghiệp, kế hoạch), (Khoa, bộ môn, cơ sở hạ tầng, hỗ trợ, hoạt động), (Đại học, sinh viên, an toàn, mô hình, giảng viên)
Cluster 3	1082	(Công nghệ, bưu chính, sinh viên, ngày hội, khen thưởng), (Học bổng, chương trình, thực tập, công nghệ, sách), (Quyết định, cán bộ, thông báo, bổ nhiệm, quy định)

Bảng 3.5: Kết quả phân cụm cấp 2 đồ thị theo chủ đề

Cụm cấp 2	Cụm cha	Số người dùng	Chủ đề
Cluster 2	1	786	(Công nghệ, chính quy, bằng tốt nghiệp, kế hoạch); (Khoa, bộ môn, cơ sở hạ tầng, hỗ trợ, hoạt động)
	2	293	(PTIT, sinh viên, an toàn, mô hình); (Khoa học, hội nghị, việc làm, nghiên cứu, giảng viên); (Công nghệ, chính quy, bằng tốt nghiệp, kế hoạch)
Cluster 3	3	1037	(Công nghệ, bưu chính, sinh viên, ngày hội, khen thưởng); (Học bổng, chương trình, thực tập, công nghệ, sách), (Quyết định, cán bộ, thông báo, bổ nhiệm, quy định)
	4	45	(Bưu chính, thông tin, thông báo, giáo dục, việc làm)

Bảng 3.4 cho thấy 3 cụm cấp 1, có hơn 1.000 người dùng. Chỉ có một chủ đề trong cụm 1. Cụm 2 và 3 có nhiều hơn ba chủ đề. Cả cụm 2 và cụm 3 được phân cụm thành nhiều hơn hai cụm phụ nhưng trong phần kết quả này chỉ giữ lại 2 cụm phụ quan trọng nhất cho sự ngắn gọn (xem Bảng 3.5). Dựa trên các kết quả phân cụm theo phân cấp được hiển thị trong Bảng 3.4 và Bảng 3.5, có thể dễ dàng chia người dùng thành 3 nhóm quan tâm lớn. Nhóm

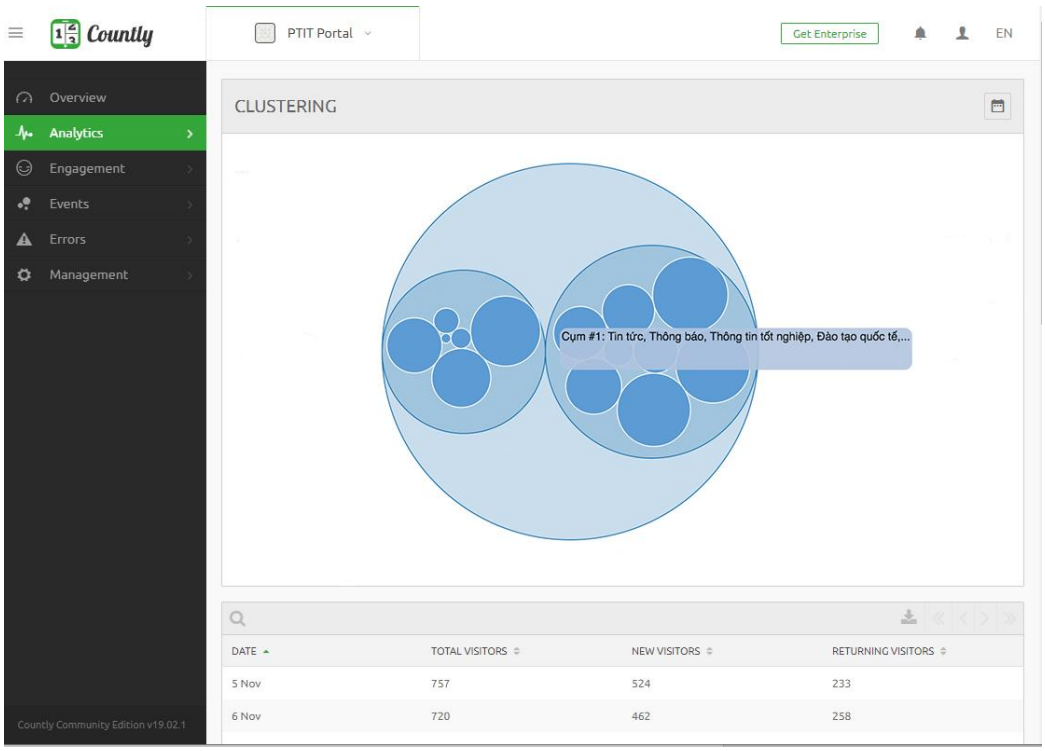
đầu tiên - nhóm lớn nhất quan tâm đến Thông báo về một số kết quả điểm chuẩn. Người dùng trong nhóm này thường là sinh viên. Kết quả này tương tự với kết quả khi phân tích đồ thị chuyên mục. Nhóm thứ hai có xu hướng thông tin của trường đại học hoặc tin tức. Một số lượng lớn người dùng trong nhóm này quan tâm đến những thứ liên quan đến chứng chỉ / văn bằng (nhóm con 1 trong Bảng 3.5) và các hoạt động trong trường đại học. Họ là những sinh viên học xong và đang chờ tốt nghiệp. Phần còn lại chú ý đến thông tin của nghiên cứu, hội nghị và trường đại học. Nhóm người dùng trong cụm 3 quan tâm nhất đến việc khen thưởng sinh viên cho một số cuộc thi và thông tin thực tập cũng như học bổng. Họ phải là những học sinh giỏi, thích những thử thách trong các cuộc thi của trường đại học. Trên thực tế, trong thời gian này, rất nhiều sinh viên trong trường đại học tham dự các cuộc thi lập trình do trường đại học và Samsung tổ chức. Một số trong số họ có thể là sinh viên năm thứ ba hoặc năm thứ tư đang tìm kiếm thông tin về chương trình thực tập hoặc học bổng từ các công ty. Có thể nhận ra rằng rất ít người dùng / sinh viên trong nhóm này quan tâm về tin tức từ trường đại học. Những phát hiện này khá giống với kết quả đã nhận được từ phân tích đồ thị chuyên mục, nhưng không có tên chuyên mục.

Đồ thị theo trang web. Áp dụng thuật toán phân cụm phân cấp vào đồ thị theo trang web, người dùng được phân thành 7 cụm. Sau đó, đối với mỗi trang web, ánh xạ tới chuyên mục và chủ đề tương ứng. Bảng 3.6 mô tả ba cụm trên cùng trong kết quả phân cụm sau khi gán tên chuyên mục. Từ kết quả, chỉ biết rằng một số lượng lớn người dùng quan tâm đến Tin tức, sau đó là Thông báo cho sinh viên, Thông báo khác và tin tức Sinh viên. Tất cả các cụm mô tả thông tin khá giống nhau. Kết quả tương tự khi gán chủ đề cho các trang web theo cụm. Lý do là nhiều trang web trong các cụm khác nhau thuộc về cùng thể loại hoặc chủ đề. Khi gán chuyên mục và chủ đề cho trang, các chuyên mục và chủ đề tương tự sẽ xuất hiện trong các trang web khác nhau. Nó dẫn đến các cụm khác nhau có thông tin tương tự.

Bảng 3.6: Kết quả phân cụm đồ thị theo trang web

Cụm	Số người dùng	Các chuyên mục
Cluster 1	5096	Tin tức, Thông báo sinh viên, Thông báo, Tin tức sinh viên
Cluster 2	184	Tin tức, Thông báo sinh viên, Thông báo, Tin tức sinh viên
Cluster 3	120	Tin tức, Thông báo, Thông báo sinh viên

3.2.4. Xây dựng giao diện công cụ phân tích log truy cập



Hình 3.3: Giao diện công cụ phân tích log truy cập website

Với quy trình thu thập và xử lý log trong thực nghiệm này, để thuận lợi cho quá trình phân tích log truy cập website và đánh giá ý nghĩa của kết quả phân tích. Do quá trình phân cụm dữ liệu này tốn nhiều thời gian để xử lý tùy thuộc vào số lượng bản ghi dữ liệu nên các tác vụ sẽ được thực hiện ở nền, quản trị viên sẽ xem các kết quả sau khi quá trình phân tích hoàn tất.

3.4. Kết luận chương

Chương 3 đã trình bày về quá trình thực nghiệm kết quả từ dữ liệu thực tế áp dụng kỹ thuật đã đề xuất ở chương 2 để đưa ra kết quả phân cụm người dùng. Kết quả phân tích trên đã phát hiện ra một số mối quan tâm của người dùng. Những kết quả này có thể cung cấp hỗ trợ đáng kể cho quản trị viên website để tối ưu hóa cấu trúc của trang web và cải thiện các chiến lược đề xuất trang web.

KẾT LUẬN VÀ KIẾN NGHỊ

Luận văn này tập trung nghiên cứu về khai phá sử dụng web, log truy cập, các kỹ thuật thu thập log truy cập website, các kỹ thuật xử lý và phân tích log. Cụ thể luận văn đã đạt được các kết quả sau:

- Nghiên cứu các khái niệm về khai phá dữ liệu, khai phá sử dụng web, tìm hiểu về quá trình khai phá sử dụng web, tổng quan về các nghiên cứu hiện nay về khai phá dữ liệu.
- Nghiên cứu các kỹ thuật thu thập log để biết được tình trạng hoạt động của các máy chủ dịch vụ, nắm bắt hành vi người dùng, giúp cải thiện các hệ thống thu thập log hiện có.
- Nghiên cứu về học không giám sát và các kỹ thuật phân cụm dữ liệu để có thể áp dụng kỹ thuật xử lý log và phân tích log truy cập website.
- Đưa ra mô hình thử nghiệm với đầy đủ các bước thu thập, chuẩn hóa, xử lý và phân tích log, có thể triển khai sử dụng trong thực tế.

Luận văn có thể phát triển tiếp theo hướng như sau:

Tiếp tục thử nghiệm với dữ liệu log một số cổng thông tin điện tử khác. Xây dựng hệ thống phân tích log truy cập website hoàn thiện, đưa ra các báo cáo trực quan, xây dựng các hệ thống gợi ý thay đổi nội dung, cấu trúc website tích hợp trực tiếp vào trang quản trị website cho các quản trị viên, … Nghiên cứu ứng dụng việc xử lý và phân tích log vào nhiều lĩnh vực khác nhau.