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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI - 2020

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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC

GS. TS. TỪ MINH PHƯƠNG

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều được trình bày hoặc là của cá nhân hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp. Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả luận văn

Nguyễn Văn Tiến

LỜI CẢM ƠN

Em xin gửi lời cảm ơn tới thầy hướng dẫn GS. TS. Từ Minh Phương, thầy đã tận tình hướng dẫn khoa học và giúp đỡ, chỉnh sửa và chỉ bảo em trong suốt quá trình nghiên cứu và hoàn thành luận văn.

Em cũng xin chân thành cảm ơn các thầy cô tại Học viện Công nghệ Bưu chính Viễn thông, đặc biệt các thầy cô khoa Công nghệ thông tin, đã tận tình dạy dỗ, giúp đỡ và tạo mọi điều kiện tốt nhất cho em trong suốt quãng thời gian em theo học tại học viện, để em có thể hoàn thành được luận văn này.

Mặc dù đã cố gắng hoàn thành luận văn nhưng chắc chắn sẽ không tránh khỏi những sai sót, em kính mong nhận được sự thông cảm và góp ý của các thầy cô và các bạn.

Luận văn này được hỗ trợ bởi Bộ Khoa học Công nghệ, thông qua đề tài mã số KC.01.23/16-20.

Em xin trân trọng cảm ơn.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
API	Application Programming Interface	Giao diện lập trình ứng dụng
CGI	Common Gateway Interface	Giao diện cổng chung
CRM	Customer relationship management	Quản lý quan hệ khách hàng
CSDL	Cơ sở dữ liệu	Cơ sở dữ liệu
HTTP	Hypertext Transfer Protocol	Giao thức truyền tải siêu văn bản
HTML	Hypertext Markup Language	Ngôn ngữ đánh dấu siêu văn bản
ISP	Internet Service Provider	Nhà cung cấp dịch vụ Internet
IP	Internet Protocol	Địa chỉ IP
PII	Personally Identifiable Information	Thông tin nhận dạng cá nhân
URI	Uniform Resource Identifier	Mã định danh tài nguyên thống nhất

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

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.

Trong thực tế, với các kỹ thuật tiên tiến hiện nay, log truy cập website được ứng dụng nhiều hơn, các kỹ thuật xử lý và phân tích log được tối ưu nhằm phục vụ cho các công việc nâng cao trải nghiệm người dùng. Thay vì ghi lại nhật ký các hoạt động của máy chủ, công việc này lại ghi các hành vi của người dùng, trên cơ sở các hành vi này, các công cụ phân tích có thể khám phá ra thói quen, sở thích của người dùng hoặc phát hiện những điểm mạnh, điểm yếu về nội dung, định dạng trang web. Từ những khám phá đó đưa ra những đề xuất thay đổi website để cải thiện trải nghiệm cho người dùng website. Luận văn này sẽ tập trung vào việc phát triển giải pháp thu thập và phân tích log truy cập website để nâng cao trải nghiệm người dùng.

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.

Một vấn đề đặt ra với kỹ thuật xây dựng hệ thống website phân tán, máy chủ website được đặt ở nhiều nơi khác nhau, việc ghi log phía máy chủ gặp nhiều vấn đề khó khăn trong việc tổng hợp để xử lý và phân tích dữ liệu. Ngoài ra ngày nay với sự

gia tăng thiết bị di động, sự phát triển của các trình duyệt máy tính. Nhiều thao tác với website được thực hiện ở máy khách (nhấp chuộn, cuộn màn hình, ...) là những thông tin có nhiều giá trị lại không ghi được log máy chủ. Thông tin ghi được chính xác về hành vi người dùng có giá trị rất lớn đến việc phân tích, đánh giá sự hoàn thiện của website nhằm nâng cao trải nghiệm của người dùng với website. Việc thu thập log truy cập người dùng từ phía máy khách có thể được lưu trữ tại một hoặc nhiều máy chủ độc lập với website vẫn có thể đáp ứng được các yêu cầu giống như ghi log ở phía máy chủ vừa có thêm nhiều thông tin hữu ích cho việc phân tích hành vi người dùng.

Chính vì vậy, 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

Chương 1 giới thiệu tổng quan về log truy cập, các khái niệm. Tìm hiểu phương pháp phân tích log truy cập bằng kỹ thuật học không giám sát.

Hai bài toán chính cần giải quyết là bài toán thu thập log truy cập website (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, …) và bài toán phân tích log truy cập website (xác định các nhóm người dùng có nhu cầu thông tin tương tự nhau dựa trên log tương tác ghi lại)

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.

Trong những năm gần đây, sự phát triển mạnh của dữ liệu lớn, các hệ thống phân tán phục vụ hàng triệu người dùng. Các hệ thống lớn như website thương mại điện tử, cổng thông tin điện tử mỗi ngày ghi nhận hàng trăm ngàn cho đến hàng triệu bản ghi log truy cập. Dựa trên các dữ liệu đã thu thập được, các nhà phát triển phải tiến hành xử lý, phân tích dữ liệu này để nắm bắt được hiện trạng thực tế của hệ thống. Bài toán phân tích log truy cập là bài toán đang được nhiều nghiên cứu quan tâm, mục tiêu của bài toán là giải quyết các vấn đề còn tồn đọng được ghi nhận (ví dụ các lỗi, các tính năng không hoạt động hoặc hoạt động chưa tốt, …) của hệ thống hiện tại để cải thiện và nâng cao chất lượng của hệ thống.

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.

Thông thường, có nhiều hình thức thu thập log truy cập. Tuy nhiên theo nhóm tác giả Jaideep Srivastava [5] và L.K. Joshila Grace [7] thu thập log website có ba hình thức phổ biến: Thu thập log ở máy chủ, thu thập log ở máy khách và thu thập log thông qua proxy.

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)

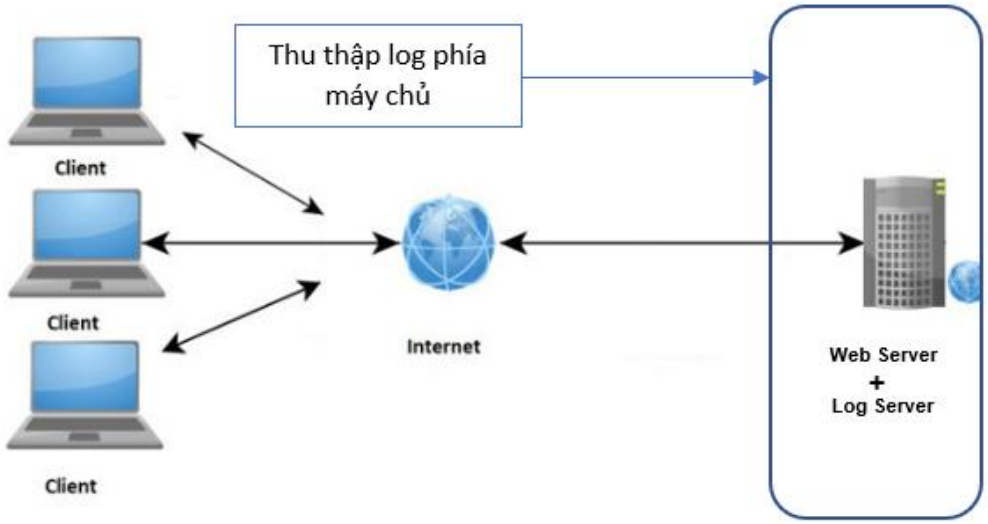
Hình 1.1: Dữ liệu log thu thập trên máy chủ

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.

Máy chủ web cũng dựa vào các tiện ích khác như tập lệnh CGI để xử lý dữ liệu được gửi lại từ trình duyệt của người dùng. Các máy chủ web triển khai tiêu chuẩn CGI phân tích URI của tệp được yêu cầu để xác định xem đó có phải là chương trình ứng dụng hay không. URI cho các chương trình CGI có thể chứa các giá trị tham số bổ sung được truyền cho ứng dụng CGI. Khi chương trình CGI đã hoàn thành việc thực thi, máy chủ Web sẽ gửi đầu ra của ứng dụng CGI trở lại trình duyệt. Hình 1.2 mô tả quá trình thu thập log phía máy chủ.

Giống như các hệ thống bình thường, phía máy chủ ứng dụng web cũng được tích hợp các công cụ để lưu lại các tác động trên hệ thống. Thông thường, các máy chủ web đều được tích hợp sẵn tính năng này. Một số được cấu hình mặc định, một số trường hợp quản trị viên phải tiến hành cấu hình các thông tin cần lưu log truy cập.

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

Tuy nhiên, giải pháp thu thập log phía máy chủ cũng có một số nhược điểm. Công nghệ web hiện nay có nhiều mức độ lưu bộ đệm ở nhiều bước khác nhau trong môi trường web. Dữ liệu bộ đệm có thể được lưu ở trình duyệt của người sử dụng, hoặc một máy chủ proxy trung gian. Ví dụ, một trang web người dùng vừa truy cập, sau đó không lâu, người dùng lại tiếp tục truy cập lại trang web này, trình duyệt có thể lấy kết quả đã được lưu trước đó để hiển thị cho người dùng. Trong trường hợp này, người dùng vẫn xem được nội dung của trang web, nhưng máy chủ hoàn toàn không biết việc người dùng đang xem trang web đó, dẫn đến dữ liệu log cũng không được ghi lại.

Có thể thấy, các giải pháp thu thập log phía máy chủ phù hợp để sử dụng trong các hệ thống website, với nhiều ưu điểm về hiệu năng, các giải pháp đều hỗ trợ các báo cáo tổng quan về hiệu năng của máy chủ, theo dõi những hoạt động bất thường

của hệ thống. Tuy nhiên các giải pháp thu thập log này không ghi lại được các tương tác của người dùng với hệ thống để giải quyết bài toán đã đưa ra.

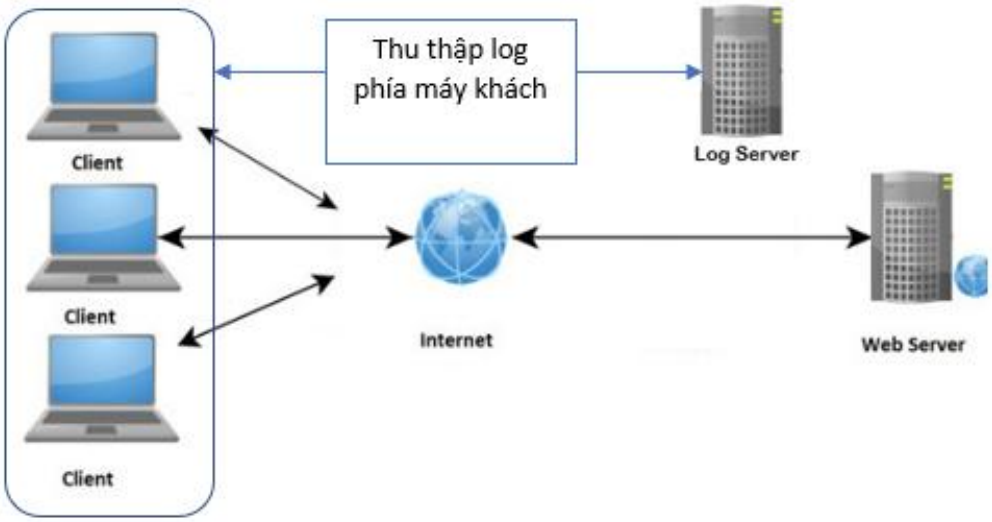
1.2.2. Phương pháp thu thập log phía máy khách

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. Thật may mắn, ngày nay, các trình duyệt phổ biến đều hỗ trợ Javascripts và mặc định được bật khi trình duyệt được cài đặt, các tính năng chạy trên công nghệ web hiện tại cũng sử dụng rất nhiều mã JavaScripts do đó phần lớn người sử dụng đều bật tính năng này để có thể trải nghiệm tốt nhất với trang web.

Ưu điểm của phương pháp này giảm tải được công việc phải xử lý cho máy chủ. Thu thập log phía máy khách giải quyết được các vấn đề liên quan đến dữ liệu được lưu trên bộ nhớ đệm ở phía máy khách hoặc qua các máy chủ proxy, hành vi duyệt web của người dùng vẫn được thu thập do mã nguồn Javascripts được thực thi trên chính trình duyệt mà người dùng sử dụng.

Thu thập log phía máy khách bao gồm 2 thành phần, mã nguồn Javascripts được thực thi tại máy khách chỉ thực hiện công việc nhận biết hành vi người dùng, có thể xử lý dữ liệu thô thành dạng dữ liệu đã được tùy biến. Đằng sau quá trình này, vẫn cần một máy chủ để lưu trữ dữ liệu phục vụ cho quá trình phân tích, khai phá dữ liệu sau này. Sau khi dữ liệu được xử lý sơ bộ ở máy khác, mã nhúng Javascripts sẽ thực hiện quá trình gửi thông tin về phía máy chủ. Máy chủ thu thập log này không nhất thiết phải đặt cùng với máy chủ web. Nó có thể hoạt động độc lập và một máy chủ thu thập log có thể thu thập dữ liệu log cho nhiều trang web thuộc nhiều máy chủ khác nhau. Hình 1.3 mô tả cách hoạt động của phương pháp thu thập log phía máy khách.

Các website trên toàn cầu ngày càng phát triển, nhu cầu thu thập log trên các webiste cũng ngày càng gia tăng, các dịch vụ thu thập log cũng được các ông lớn trong làng công nghệ chú trọng phát triển. Các công cụ thu thập log được xây dựng sẵn để dễ dàng triển khai, tùy vào tính chất và tính năng các công cụ này có thể miễn phí hoặc trả phí.



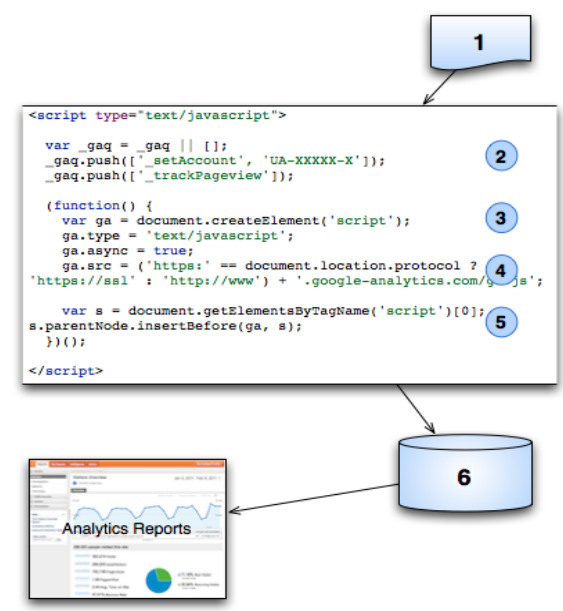
Hình 1.3: Mô hình thu thập log phía máy khách

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.



Hình 1.4: Mô hình hoạt động của Google Analytics

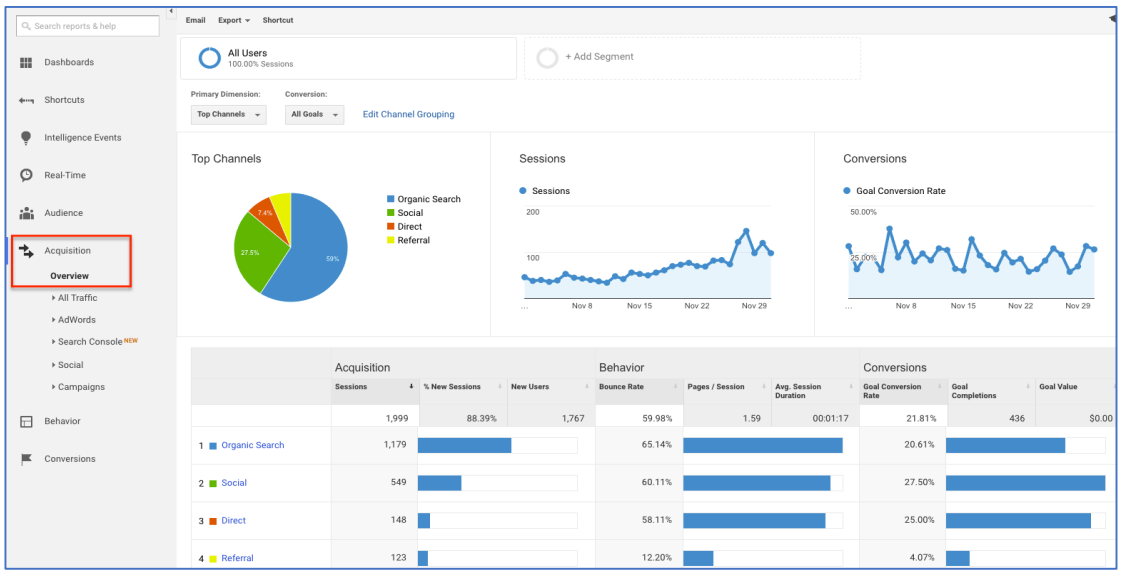
Quyền riêng tư đã trở thành một chủ đề quan trọng trong phân tích trang web. Là một vấn đề thực tiễn tốt nhất, không nên theo dõi Thông tin nhận dạng cá nhân (Personally Identifiable Information - PII). Dưới đây là tóm tắt nhanh về thông tin liên quan đến quyền riêng tư liên quan đến Google Analytics:

- Google Analytics không báo cáo về thông tin nhận dạng cá nhân.
- Cookie Google Analytics thu thập dữ liệu nhật ký Internet tiêu chuẩn theo cách ẩn danh và Google không chia sẻ dữ liệu này với bất kỳ bên thứ ba nào.
- Google Analytics không theo dõi người dùng trên nhiều trang web không liên quan.
- Quản trị viên có thể chọn chia sẻ dữ liệu của mình ẩn danh với Google hoặc có thể từ chối dịch vụ này.
- Google cung cấp tiện ích trên trình duyệt cho phép người dùng hoàn toàn từ chối theo dõi Google Analytics.

- Google cũng cung cấp một phương pháp gọi là Ẩn danh IP, cú pháp: _anonymousizeIp(), để xáo trộn thông tin IP được gửi tới Google. Điều này ngăn Google Analytics báo cáo thông tin vị trí địa lý.

Google Analytics cung cấp nhiều dữ liệu về lưu lượng truy cập trang web, nhưng phân tích chính xác về Google Analytics có thể cho quản trị biết nhiều hơn chỉ đơn giản là lưu lượng truy cập mà trang web đang nhận được.

Quản trị viên có thể tìm hiểu tất cả các loại thông tin về những thứ như nguồn lưu lượng truy cập, ví dụ: trang web nào đã giới thiệu lưu lượng truy cập đến trang web, kênh truyền thông xã hội nào đang đưa khách truy cập trực tiếp đến trang web, từ khóa nào được xếp hạng trên Google hay các công cụ tìm kiếm khác.



Hình 1.5: Giao diện công cụ Google Analytics

Google Analytics cũng đo lường các loại lưu lượng khác nhau để xác định loại nào có giá trị hơn, ví dụ như lưu lượng truy cập công cụ tìm kiếm so với lưu lượng phương tiện truyền thông xã hội.

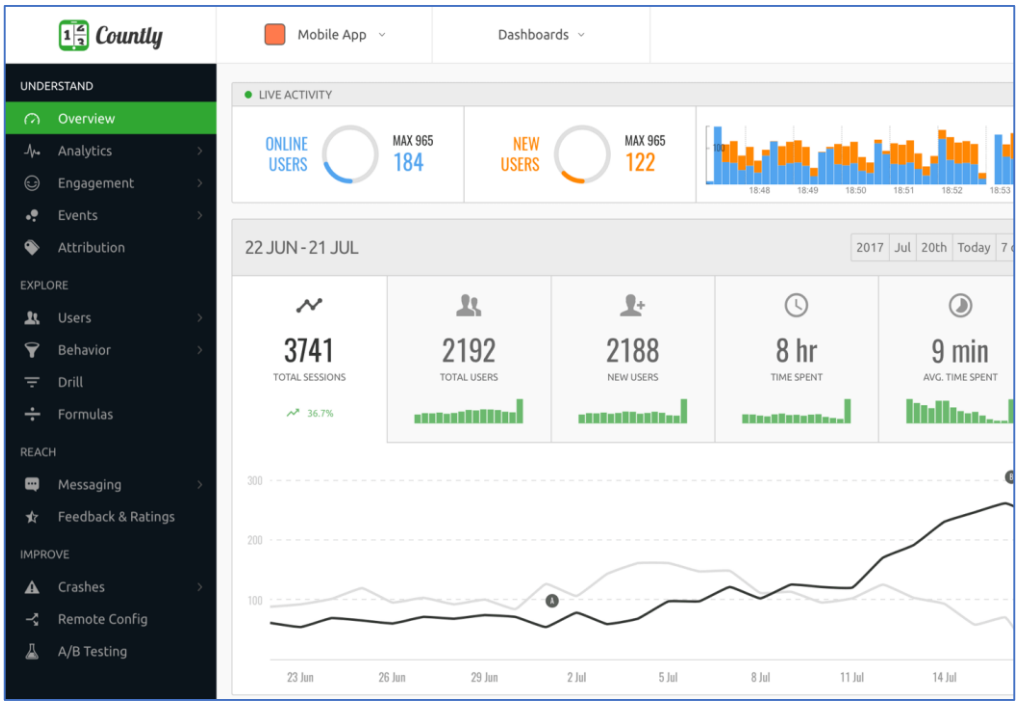
Google Analytics có thể giúp tối ưu hóa thị trường, dẫn đến tăng trưởng doanh thu cho hoạt động kinh doanh.

Hình 1.6: Thống kê theo vị trí địa lý của người dùng của Google Analytics

Google Analytics giúp người quản trị dễ dàng hiểu cách người dùng trang web và ứng dụng tương tác với nội dung của trang web, vì vậy quản trị viên biết những gì mà người dùng hoạt động trên website và những gì người dùng không quan tâm. Xem cách mọi người tương tác với các trang web và ứng dụng của và vai trò của các kênh khác nhau bằng cách xem các báo cáo và bảng điều khiển. Người quản trị trang web thậm chí có thể kết nối các hệ thống được sử dụng để đo lường CRM, điểm bán hàng và các điểm tiếp xúc khác với khách hàng để có cái nhìn đầy đủ hơn.

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.



Hình 1.7: Công cụ thu thập log Countly

Countly là một phần mềm phân tích trang web nguồn mở, miễn phí cho phép người quản trị xem những gì người dùng đang làm trên trang web của họ.

Countly hỗ trợ các sự kiện có thể được ghi vào cơ sở dữ liệu một cách không đồng bộ và ghi nhật ký sự kiện và giao diện người dùng quản trị / báo cáo có thể chạy trên cùng một máy chủ hoặc riêng biệt. Countly có thể chạy trên nhiều máy chủ web phân tán và ghi vào cơ sở dữ liệu từ xa.

Countly ra đời vì nhu cầu về một khung mã nguồn mở có thể được sử dụng để dễ dàng thêm các tính năng phân tích trang web vào các trang web và ứng dụng. Countly có thể theo dõi và phân tích cách mọi người sử dụng các trang web và ứng dụng. Countly được cấp phép theo GPL và cung cấp cho chủ sở hữu và nhà phát triển trang web những cách dễ dàng để thêm phân tích trang web vào trang web của họ bằng các API dựa trên Javascript, NodeJS hoặc REST đơn giản. Countly cũng hỗ trợ tích hợp để theo dõi các trang web được tạo bằng các khung quản lý nội dung phổ biến như WordPress và MediaWiki. Countly có Plugin để xác thực, đối tượng truy cập cơ sở dữ liệu, xác thực dữ liệu và vị trí địa lý.

Hình 1.8: Thống kê theo vị trí địa lý của người dùng của Countly

Điểm mạnh là với Countly, người quản trị có quyền sở hữu dữ liệu của mình. Nó lưu trữ dữ liệu trong cơ sở dữ liệu và người quản trị quyết định ai sẽ chia sẻ dữ liệu đó và trong bao lâu mình muốn lưu trữ dữ liệu đó. Nói cách khác, người quản trị tránh được Google hoặc bất kỳ công ty nào khác có mối quan tâm tiềm năng đối với hành vi của người dùng trên trang web khỏi dữ liệu phân tích trang web của mình.

Quản trị viên có quyền kiểm soát hoàn toàn, thiết lập máy chủ của riêng và dữ liệu riêng. Các dữ liệu được thu thập và lưu trữ cũng có nhiều thông tin hơn so với Google Analytics do chính sách riêng tư của Google. Tuy nhiên, đối với người mới làm quen phân tích log truy cập trang web, điều này có thể là một bất lợi do phải tự cài đặt hệ thống, xử lý các lỗi phát sinh. Ngoài ra Countly cần cài đặt trên máy chủ riêng nên sẽ phát sinh các chi phí về duy trì máy chủ.

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.

Nếu đang sử dụng máy chủ proxy, lưu lượng truy cập Internet sẽ truyền qua máy chủ proxy theo đường của nó đến địa chỉ của máy chủ. Sau đó, yêu cầu này sẽ trở lại cùng một máy chủ proxy (cũng xảy ra trường hợp ngoại lệ đối với quy tắc này) và máy chủ proxy đó sẽ chuyển tiếp dữ liệu nhận được từ website đến người dùng.

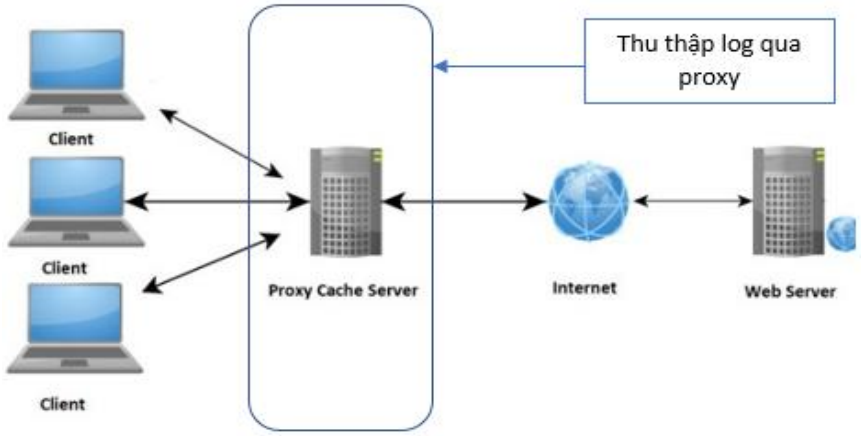
Các máy chủ proxy hiện đại thực hiện nhiều công việc hơn ngoài việc chuyển tiếp các yêu cầu web, nó còn thực hiện bảo mật dữ liệu và tăng hiệu suất mạng. Các máy chủ proxy hoạt động như tường lửa và bộ lọc web, cung cấp kết nối mạng chia sẻ và dữ liệu bộ nhớ cache để tăng tốc các yêu cầu thông thường. Một máy chủ proxy tốt sẽ bảo vệ người dùng và mạng nội bộ khỏi các thứ không mong muốn từ Internet. Cuối cùng, máy chủ proxy có thể cung cấp mức độ riêng tư cao.

Caching của các trang web có thể cải thiện chất lượng dịch vụ của một mạng theo 3 cách. Thứ nhất, nó có thể bảo tồn băng thông mạng, tăng khả năng mở rộng. Tiếp đến, có thể cải thiện khả năng đáp trả cho các máy khách. Ví dụ, với một bộ đệm HTTP, Trang web có thể tải nhanh hơn trong trình duyệt web. Cuối cùng, các máy chủ proxy có thể tăng khả năng phục vụ. Các trang web hoặc các dòng khác trong bộ đệm vẫn còn khả năng truy cập thậm chí nguồn nguyên bản hoặc liên kết mạng trung gian bị ngoại tuyến.

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 1.9 cho thấy cách hoạt động của phương pháp thu thập log thông qua proxy.

Vì proxy là một máy chủ, nên việc thực hiện thu thập log cũng tương tự với các giải pháp thu thập log phía máy chủ. Ngoài ra, Việc xây dựng máy chủ proxy chủ yếu do các doanh nghiệp lớn thực hiện, hoặc proxy của các ISP, do đó các giải pháp thu thập, phân tích log chuyên nghiệp thường được cung cấp bởi các các công ty chuyên cung cấp các giải pháp cho doanh nghiệp.



Hình 1.9: 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. Bảng 1.1 cho thấy ưu, nhược điểm giữa các giải pháp.

Bảng 1.1: Ưu, nhược điểm của các giải pháp thu thập log

Giải pháp	Ưu điểm	Nhược điểm
Thu thập log phía máy chủ	Đơn giản, không cần cài đặt thêm nhiều phần mềm	Không lấy được toàn bộ hành vi người dùng do có nhiều bộ đệm trong môi trường web

Thu thập log phía máy khách	Lấy được chính xác về hành vi người dùng trên trang web	Phải cài đặt thêm các công cụ, phần mềm hỗ trợ, thêm máy chủ lưu trữ mới có thể làm chủ dữ liệu
Thu thập log qua Proxy	Hữu ích cho việc cải thiện hiệu năng ở các hệ thống mạng nội bộ	Không lấy được chính xác hành vi người dùng, chỉ lấy được log khi người dùng sử dụng proxy

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.

Một cách toán học, Học không giám sát là kỹ thuật học khi chúng ta chỉ có dữ liệu vào X mà không biết nhãn Y tương ứng.

Những thuật toán loại này được gọi là Học không giám sát vì chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào, không có câu trả lời đúng và không có vị "giáo viên" nào cả. Các thuật toán được tạo ra chỉ để khám phá và thể hiện các

cấu trúc hữu ích bên trong dữ liệu. Cụm từ không giám sát được đặt tên theo nghĩa này.

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

- Phân cụm (clustering): Một bài toán phân cụm toàn bộ dữ liệu X thành các cụm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ: phân cụm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

- Học luật kết hợp (association rule mining): Là bài toán khi chúng ta muốn khám phá ra một quy luật dựa trên nhiều dữ liệu cho trước. Ví dụ: những khách hàng nam mua quần áo thường có xu hướng mua thêm đồng hồ hoặc thắt lưng; những khán giả xem phim Spider Man thường có xu hướng xem thêm phim Bat Man, dựa vào đó tạo ra một hệ thống gợi ý khách hàng (Recommendation System), thúc đẩy nhu cầu mua sắm.

1.3.2. Một số kỹ thuật phân cụm dữ liệu

Mục đích chính của phân cụm dữ liệu nhằm khám phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó nó cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khám phá và tìm kiếm thông tin tiềm ẩn, hữu ích phục vụ cho việc ra quyết định. Ví dụ: Nhóm sinh viên trong CSDL của một trường Đại học có khả năng sắp tốt nghiệp. Như vậy, Phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và nó phổ biến, nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm.

Ta có thể khái quát hóa khái niệm Phân cụm dữ liệu [1]: 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.

Độ tương tự được xác định dựa trên các giá trị của thuộc tính mô tả đối tượng. Thông thường, phép đo khoảng cách thường được sử dụng để đánh giá độ tương tự.

Trong học máy, Phân cụm dữ liệu được coi là thuật toán học không giám sát, vì nó phải giải quyết vấn đề tìm một cấu trúc trong tập hợp dữ liệu chưa biết trước các thông tin về lớp hay các thông tin về tập huấn luyện.

Phân cụm dữ liệu là một bài toán khó vì người ta phải giải quyết các vấn đề con như sau:

- Biểu diễn dữ liệu.

- Xây dựng hàm tính độ tương tự.

- Xây dựng các tiêu chuẩn phân cụm.

- Xây dựng mô hình cho cấu trúc cụm dữ liệu.

- Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo.

- Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm.

Theo các nghiên cứu thì đến nay chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc dữ liệu. Hơn nữa, các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc cụm dữ liệu khác nhau, với mỗi cách thức biểu diễn khác nhau sẽ có một thuật toán phân cụm phù hợp. 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

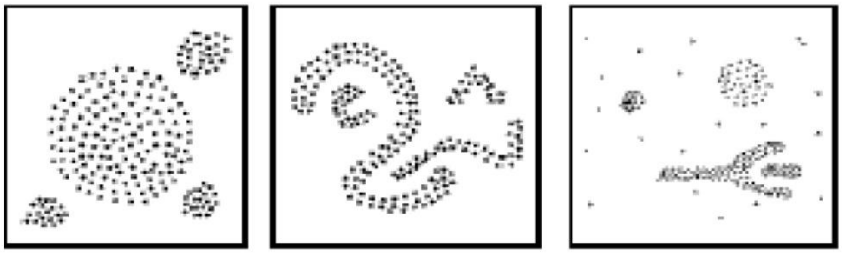
Phân cụm phân hoạch (partitioning) với ý tưởng chính là phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Các thuật toán phân hoạch có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, vì nó phải tìm kiếm tất cả các cách phân hoạch có thể được.

Chính vì vậy, trên thực tế người ta thường đi tìm giải pháp tối ưu cục bộ cho các vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của các cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Với chiến lược này, thông thường người ta bắt đầu khởi tạo một phân hoạch ban đầu cho tập dữ liệu theo phép ngẫu nhiên hoặc heuristic và liên tục tinh chỉnh nó cho đến khi thu được một phân hoạch mong muốn, thỏa mãn các điều kiện ràng buộc cho trước. Các thuật toán phân cụm phân hoạch cố gắng cải tiến tiêu chuẩn phân cụm bằng các tính các giá trị độ đo tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu. Như vậy ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược tham lam để tìm kiếm nghiệm.

1.3.2.2. Phân cụm theo mật độ

Phương pháp phân cụm này nhóm các đối tượng theo hàm mật độ xác định. Mật độ được định nghĩa như là số các đối tượng lân cận của một đối tượng dữ liệu theo một ngưỡng nào đó. Trong cách tiếp cận này, khi một cụm dữ liệu đã xác định thì nó tiếp tục được. phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận của các đối tượng này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa vào mật độ của các đối tượng để xác định các cụm dữ liệu và có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Tuy vậy, việc xác định các tham số mật độ của thuật toán là rất khó khăn, trong khi các tham số này lại có

tác động rất lớn đến kết quả Phân cụm dữ liệu. Hình minh họa về các cụm dữ liệu với các hình thù khác nhau dựa trên mật độ được khám phá từ 3 CSDL khác nhau:



Hình 1.10: Một số dạng khám phá bởi phân cụm dựa trên mật độ

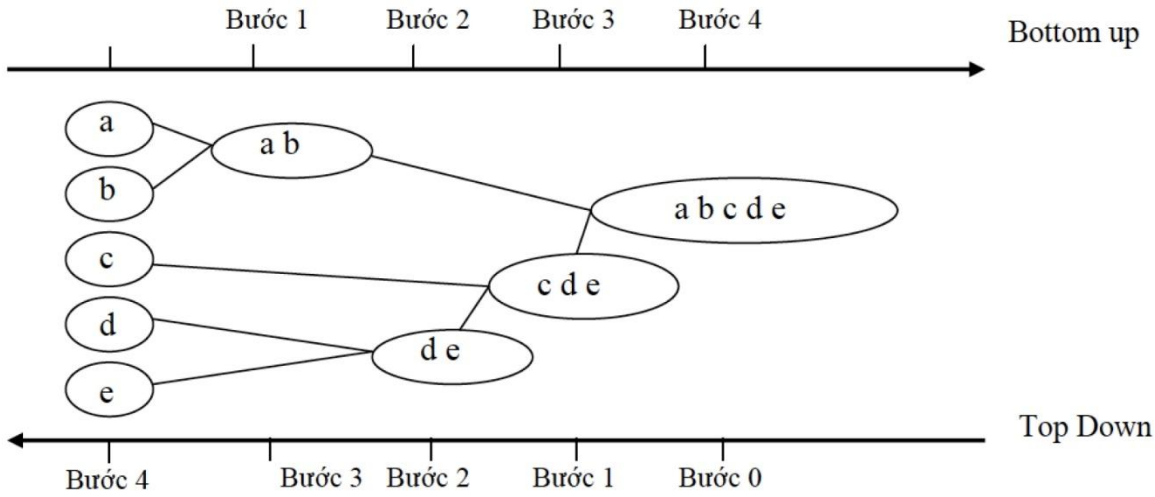
Các cụm có thể được xem như các vùng có mật độ cao, được tách ra bởi các vùng không có hoặc có mật độ thấp, khái niệm mật độ ở đây được xem như là các số các đối tượng lân cận.

1.3.2.3. Phân cụm phân cấp

Phân cụm phân cấp sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Cây phân cụm có thể được xây dựng theo hai phương pháp tổng quát là: Trên xuống (Top down) và phương pháp Dưới lên (Bottom up).

Phương pháp Top down: Bắt đầu với trạng thái là tất cả các đối tượng được xếp trong cùng một cụm. Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự nào đó cho đến kho mỗi đối tượng là một cụm hoặc cho đến khi điều kiện dừng thỏa mãn. Cách tiếp cận này sử dụng chiến lược chia để trị trọng quá trình phân cụm.

Phương pháp Bottom up: Phương pháp này bắt đầu với mỗi đối tượng được khởi tạo tương ứng với các cụm riêng biệt, sau đó tiến hành nhóm các đối tượng với nhau theo một độ đo tương tự, quá trình này được thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm hoặc cho đến khi các điều kiện dừng thỏa mãn. Như vậy, các tiếp cận này sử dụng chiến lược tham lam trong quá trình phân cụm.



Hình 1.11: Các chiến lược phân cụm phân cấp

Trong thực tế áp dụng, có nhiều trường hợp người ta kết hợp cả hai phương pháp phân cụm phân hoạch và phương pháp phân cụm phân cấp, nghĩa là kết quả thu được của phương pháp phân cấp có thể cải tiến thoogn qua bước phân cụm phân hoạch. Phân cụm phân hoạch và phân cụm phân cấp là hai phương pháp phân cụm dữ liệu cổ điển, hiện nay đã có nhiều thuật toán cải tiến dựa trên hai phương pháp này được áp dụng phổ biến trong Khai phá dữ liệu.

Với bài toán xác định các nhóm người dùng có nhu cầu thông tin tương tự nhau, kỹ thuật phân cụm phân cấp phù hợp và đơn giản, với kỹ thuật này, có thể chia tập hợp người dùng ban đầu thành các nhóm có chiều sâu. Ví dụ, nhóm người dùng sinh viên, có thể chứa các nhóm người dùng sinh viên năm nhất, sinh viên năm cuối,…

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 - PHƯƠNG PHÁP THU THẬP VÀ PHÂN TÍCH LOG TRUY CẬP WEBSITE

Chương 2 trình bày cụ thể về phương pháp xây dựng giải pháp thu thập log truy cập website, giải pháp phân tích log truy cập webiste dựa trên kỹ thuật phân cụm, cách xác định độ tương tự giữa hai người dùng, tìm hiểu về thuật toán phân cụm và cách xác định ý nghĩa của các cụm.

2.1. Xây dựng công cụ thu thập log

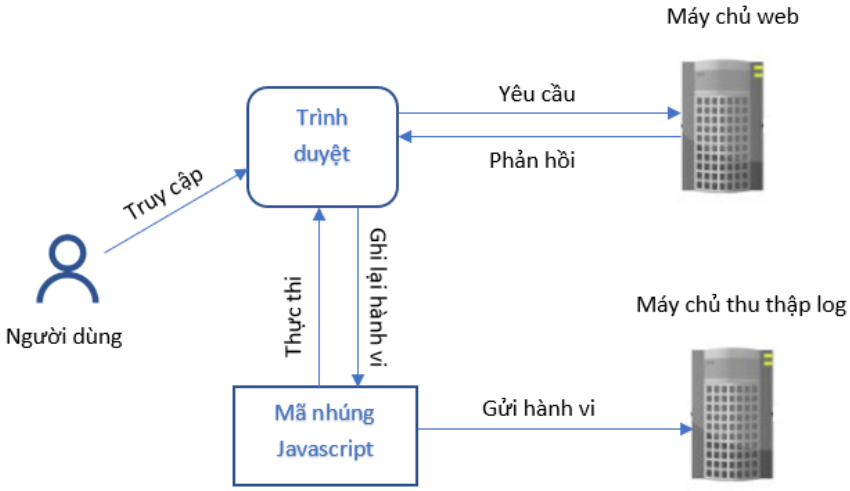
Ngày nay, các công cụ phân tích website được cải tiến không ngừng. Nó hỗ trợ cho người quản trị website có thể nắm được các số liệu thống kê, phân tích về website của mình. Một số công cụ còn dựa vào cookies, thông tin của trình duyệt, kết hợp với kho dữ liệu khổng lồ của họ để xác định độ tuổi, giới 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.

Tuy nhiên, các công cụ này được xây dựng sẵn, người quản trị không thể làm chủ dữ liệu của mình, và buộc phải chia sẻ dữ liệu cho bên thứ ba. Nhằm giải quyết vấn đề này, chúng ta có thể xây dựng công cụ riêng để thống kê, phân tích dữ liệu từ log truy cập website đã thu thập được bằng cách áp dụng kỹ thuật học không giám sát. Trong luận văn này, ta sẽ xem xét cách áp dụng phương pháp phân cụm phân cấp để phân cụm người dùng website từ dữ liệu log thu thập được.

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.

Thông thường, người dùng website không nhất thiết phải đăng nhập hay khai báo bất kỳ thông tin cá nhân nào, đặc biệt với các website tin tức, báo chí, ... Do đó, trong CSDL người dùng của website không có bất kỳ thông tin cá nhân nào của người dùng, thậm chí một người dùng truy cập website vào thời điểm khác nhau, cũng khó khăn để xác định các phiên truy cập đó là cùng một người. Việc này đòi hỏi xây dựng

một công cụ để thu thập log để xác định được một số thông tin như địa chỉ IP, loại trình duyệt, cookies và một số dấu hiệu khác từ người dùng để phân biệt các người dùng duyệt web trong hệ thống một cách chính xác.



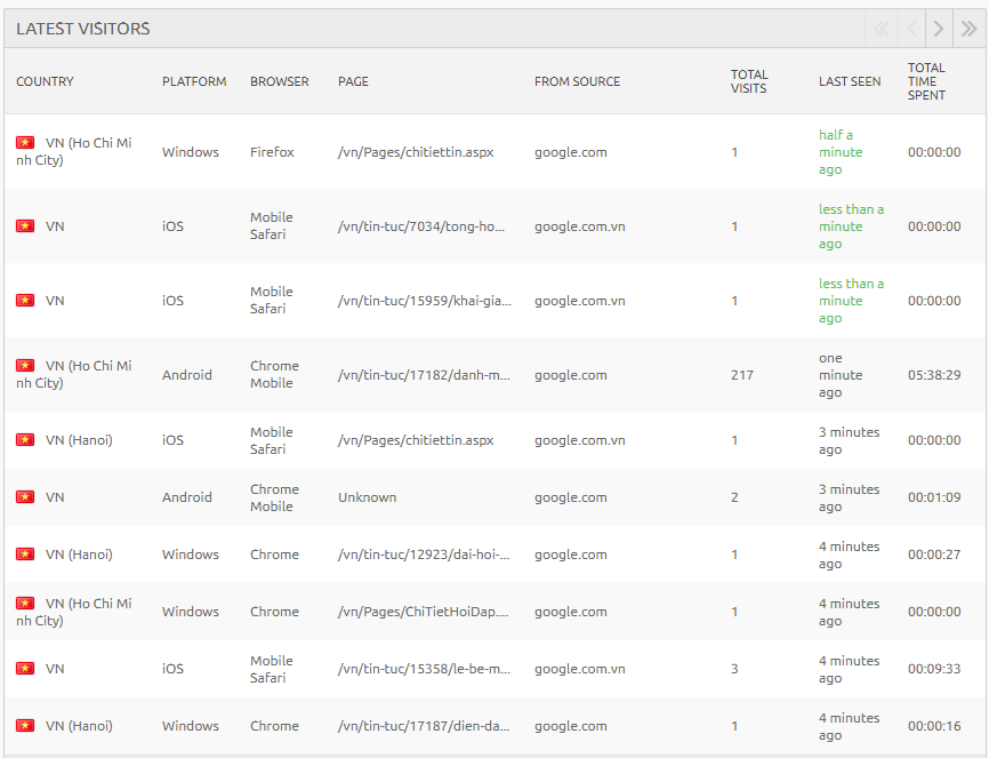
Hình 2.1: Sơ đồ mô tả hoạt động hệ thống thu thập log

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. Vai trò của các thành phần như sau:

- Trình duyệt: Khi có người dùng truy cập, trình duyệt gửi yêu cầu đến máy chủ web.
- Máy chủ web: Phản hồi khi có yêu cầu ghé thăm trang web của người dùng. Mã phản hồi đã được tích hợp mã nhúng Javascript và gửi đến trình duyệt.
- Mã nhúng Javascript: Thực thi trên trình duyệt, ghi nhận lại các hành vi tương tác của người dùng với website sau đó gửi các thông tin về hành vi này cho máy chủ thu thập log.
- Máy chủ thu thập log: Khi nhận được thông tin hành vi của người dùng, tiến hành xử lý các thông tin và lưu trữ thông tin vào CSDL.

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.



LATEST VISITORS

COUNTRY	PLATFORM	BROWSER	PAGE	FROM SOURCE	TOTAL VISITS	LAST SEEN	TOTAL TIME SPENT
VN (Ho Chi Minh City)	Windows	Firefox	/vn/Pages/chitiettin.aspx	google.com	1	half a minute ago	00:00:00
VN	iOS	Mobile Safari	/vn/tin-tuc/7034/tong-ho...	google.com.vn	1	less than a minute ago	00:00:00
VN	iOS	Mobile Safari	/vn/tin-tuc/15959/khai-gia...	google.com.vn	1	less than a minute ago	00:00:00
VN (Ho Chi Minh City)	Android	Chrome Mobile	/vn/tin-tuc/17182/danh-m...	google.com	217	one minute ago	05:38:29
VN (Hanoi)	iOS	Mobile Safari	/vn/Pages/chitiettin.aspx	google.com.vn	1	3 minutes ago	00:00:00
VN	Android	Chrome Mobile	Unknown	google.com	2	3 minutes ago	00:01:09
VN (Hanoi)	Windows	Chrome	/vn/tin-tuc/12923/dai-hoi-...	google.com	1	4 minutes ago	00:00:27
VN (Ho Chi Minh City)	Windows	Chrome	/vn/Pages/ChiTietHoiDap....	google.com	1	4 minutes ago	00:00:00
VN	iOS	Mobile Safari	/vn/tin-tuc/15358/le-be-m...	google.com.vn	3	4 minutes ago	00:09:33
VN (Hanoi)	Windows	Chrome	/vn/tin-tuc/17187/dien-da...	google.com	1	4 minutes ago	00:00:16

Hình 2.2: Log truy cập thu thập được trong 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.

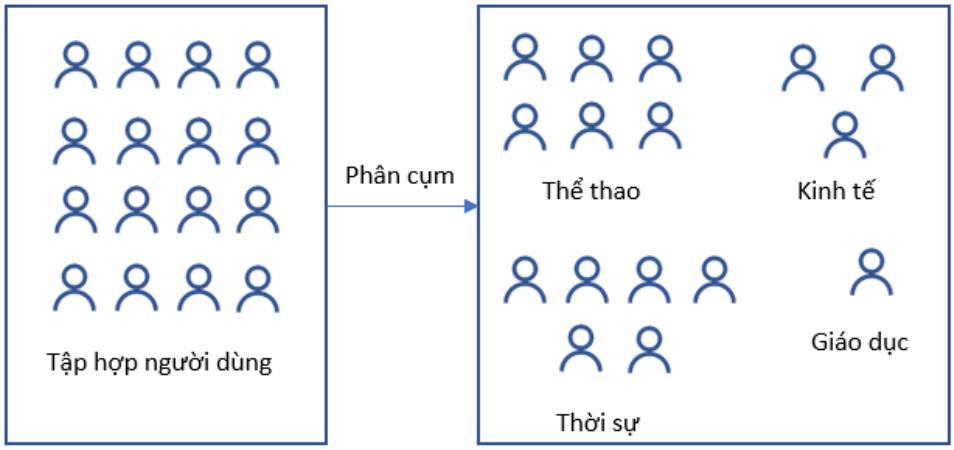
Trong phạm vi luận văn, hai người dùng được coi là có sở thích giống nhau nếu cùng xem các thông tin giống nhau. Thông tin được xác định ở các mức khác nhau. Cụ thể, hai người dùng được coi là tương tự nếu:

a. Cùng xem những trang Web giống nhau

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

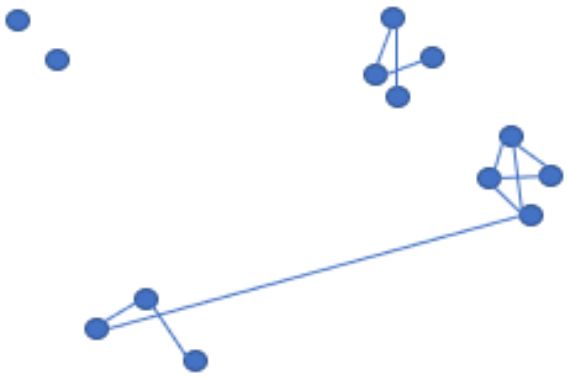
c. Cùng xem những trang Web về các chủ để giống nhau

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 dưới đây minh họa cho quá trình phân cụm người dùng.



Hình 2.3: 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 2.4: 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

Trước tiên, tiến hành loại bỏ các bản ghi dư thừa bằng cách bỏ đi những bản ghi không có giá trị đối với quá trình phân cụm người dùng. Ví dụ như cách xác định hai người dùng tương tự nhau dựa trên các hành vi cuộn chuột, nhấp chuột thì cần quan tâm đến thứ tự, thời gian xảy ra các hành vi của người dùng, còn đối với cách xác định hai người dùng tương tự dựa trên việc ghé thăm các trang web tương tự nhau thì các thông tin về nhấp chuột, thao tác chuột hay thời gian có thể không cần thiết, loại bỏ các thông tin dư thừa này sẽ giúp quá trình xử lý dữ liệu giảm bớt được thời gian tính toán đáng kể.

Ngoài ra các bản ghi liên tiếp giống nhau của cùng một người dùng trong một thời gian ngắn cũng có thể được loại bỏ. Ví dụ trong 1 giây liên tiếp, người dùng thực hiện click chuột vào các vị trí gần nhau không xác định ảnh, đường dẫn cụ thể, hoặc trong 1 giây, có 2 lượt xem trang giống nhau của cùng một người dùng. Bảng sau cho thấy dung lượng của dữ liệu sau khi xử lý.

Bảng 2.1: Loại bỏ dữ liệu dư thừa

Giá trị	Dữ liệu ban đầu	Dữ liệu sau khi loại bỏ
Số bản ghi lượt xem	153,085	136,521
Số bản ghi sự kiện chuột	461,041	210,092
Số lượng bản ghi tìm kiếm	51,112	45,017

Tiếp theo, cần chuẩn hóa các địa chỉ trang web. Ví dụ: trang web /thong-bao/1?ref=fb và /thong-bao/1 có thể cùng là một nội dung người dùng quan tâm. Trường hợp này xảy ra tùy thuộc vào cách thiết kế đường dẫn website của trang web, không phải lúc nào các đường dẫn có tiền tố giống nhau cũng thể hiện cùng nội dung. Ví dụ với trang /thong-bao.php?id=1 và /thong-bao.php?id=2 là hai trang web có nội dung khác nhau

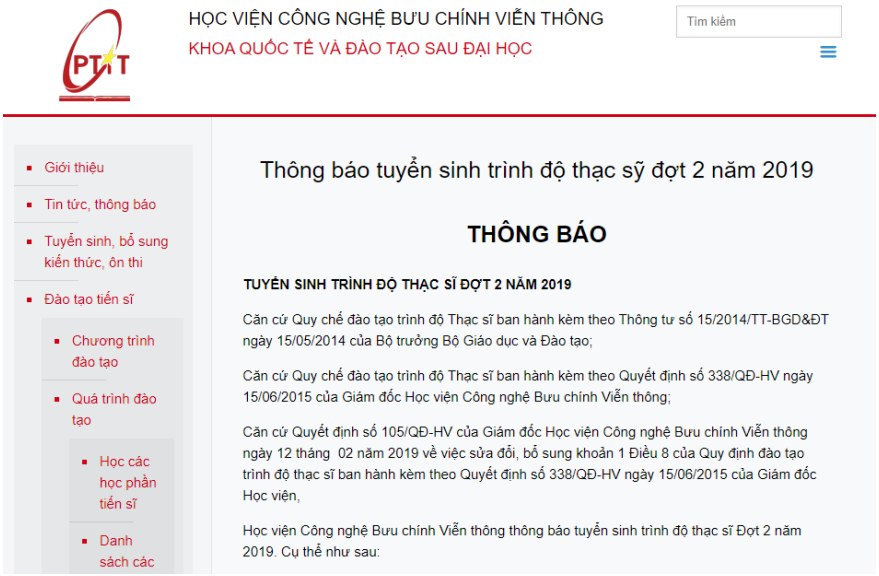
cho hai sản phẩm riêng biệt. Tùy vào hệ thống, tùy vào website mà cần xây dựng phương pháp chuẩn hóa riêng để có dữ liệu phù hợp cho quá trình phân cụm người dùng.

Một số trang web sử dụng cả 2 giao thức HTTP và HTTPS cho nội dung website của mình, thậm chí, cùng một website có nhiều tên miền khác nhau, để có kết quả phân tích dữ liệu chính xác hơn, cần xem xét các giao thức truy cập website và các tên miền thống nhất.

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, ... Bằng một số kỹ thuật, ta có thể xây dựng chương trình riêng để ánh xạ các địa chỉ trang web sang một nhóm các chuyên mục. 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 các trang web mà mỗi trang web con được chia theo các chuyên mục cố định trước. Ta có thể dựa vào truy vấn CSDL để xác định các chuyên mục của các trang web.



Hình 2.5: Trang web được xác định các chuyên mục trước

Ngược lại, đối với những trang web không được chia các chuyên mục cố định, ta có thể sử dụng thuật toán LDA (Latent Dirichlet Allocation) [3] để xác định các chủ đề cho mỗi trang web.

Thuật toán LDA là một trong những phương pháp Topic Modeling được sử dụng nhiều nhất. LDA miêu tả các văn bản như là sự pha trộn của các topics (bao gồm các từ * trọng số của các từ đó) với các xác suất nhất định. Các phân bố topic trong LDA được giả định theo phân bố Dirichlet thưa (hay sparse Dirichlet), với mục đích biểu thị rằng các đoạn văn bản (document) được biểu diễn bằng 1 số các chủ đề và các chủ đề đó lại được biểu diễn bằng 1 tập nhỏ các từ (với trọng số ứng với từng từ giảm dần).

Website được cấu thành từ nhiều trang web nhỏ (web page). Mỗi trang web này có nội dung khác nhau, người dùng quan tâm đến từng nội dung của trang web bằng cách đọc nội dung các trang web này. Khi xác định chủ đề của các trang web dựa vào tiêu đề hoặc nội dung của trang web này, sử dụng LDA sẽ xác định được các trang web thuộc các chủ đề khác nhau. Các chủ đề này được xác định theo số lượng cho trước.

Ví dụ dưới đây, sử dụng thuật toán LDA để xác định 10 chủ đề theo nội dung của các trang web. Các chủ đề bao gồm nhiều từ khóa khác nhau có xác suất xuất hiện nhỏ dần.

Bảng 2.2: Xác định các chủ đề với LDA

STT	Các từ khóa chủ chủ đề
1	viet_nam, trung_tam, samsung, tuyen_dung, svmc, ky_su, lam_viec, cong_tac, tham_du, van_hanh
2	hoc_bong, chuong_trinh, thuc_tap, hoi_thao, lap_trinh, hoc_sinh, han_quoc, thong_bao, nhung, cach_mang
3	đai_hoc, chinh_quy, tuyen_sinh, uu_tien, cao_đang, du_an, sinh_vien, linh_vuc, vien, pham_đuc_huy
4	tot_nghiep, ket_qua, phan_mem, quoc_te, to_chuc, tong_ket, lich, tieu_chuan, chang, tieu_bieu
5	thong_bao, đai_hoc, trao_đoi, đau_ra, thuc_tap_sinh, giao_duc, trung_tuyen, khu_vuc, soi_noi, thuong_mai

6	cong_nghe, vien_thong, buu_chinh, giac, suat, nhung, marketing, trung_bay, singapore, chung_ket
7	nam_hoc, hoc_phi, hoc_ky, ho_so, mien, ke_hoach, chinh_quy, van_bang, chi_phi, hoc_tap
8	sinh_vien, khoi_nghiep, sang_tao, cuoc, y_tuong, thi_sinh, tham_gia, nhan_luc, thong_minh, giai
9	quyet_đinh, cong_bo, can_bo, cong_nghe, bo_nhiem, hoc_vien, chung_ket, sinh_thai, huong_dan, trien_khai
10	đao_tao, khoa_hoc, co_so, truong, hoc_vien, phuong_thuc, sydney, thac_sy, hoan_phat, chuan

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.

Để tối ưu cho quá trình xây dựng đồ thị tương tự, từ dữ liệu log đã được xử lý, ta xây dựng một đồ thị khác, làm trung gian cho quá trình tính toán đồ thị tương tự. Đồ thị này thể hiện sự liên quan giữa người dùng và hành vi trên website, cụ thể ở đây là hành vi xem trang web.

Ta đánh số thứ tự cho người dùng trong hệ thống. Gọi N_{page} là tổng số trang web, N_{user} là tổng số người dùng web. Việc làm này để thuận tiện cho biểu diễn đồ thị ma trận kề. Ví dụ được thể hiện như bảng sau.

Bảng 2.3: Đánh số thứ tự cho người dùng truy cập

Mã người dùng	Số thứ tự
c86d75c4-9d34-4de5-9250-07843b9ae2f4	1
93838e69-8fe5-4bc5-8dfc-f09065b11244	2
72d9873e-e4bd-46ab-b99b-ff6f0d706d67	3
51ea4ff4-d043-4553-a27c-649d81b9b616	4
ad8db423-0ab3-4f4b-b7fb-1511a2319911	5
…	…
8c5defd7-a0c9-493e-8650-5509cff3b283	N_{user}

Tương tự, ta đánh số thứ tự cho các đường dẫn trang web đã được xem. Một ví dụ đơn giản như bảng dưới đây:

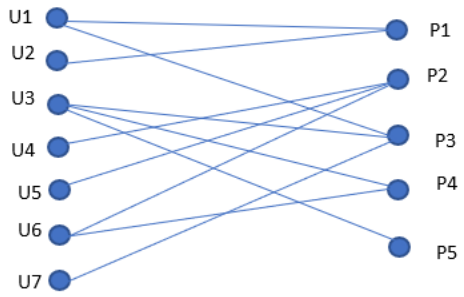
Bảng 2.4: Đánh số thứ tự cho đường dẫn trang web

Trang web	Số thứ tự
/tin-tuc/1	1
/tin-tuc/2	2
/thong-bao/1	3
/giao-duc/1	4
/kinh-te/1	5
…	…
/thoi-su/1	N_{page}

Sau khi đánh số thứ tự cho người dùng và đường dẫn trang web, dễ dàng xây dựng được đồ thị trung gian là số lượt xem trang web của mỗi người dùng được xác định như sau.

$$P[u,p] = \begin{cases} 1, & \text{nếu } u \text{ truy cập } p \\ 0, & \text{ngược lại} \end{cases}$$

$$với\ \ 0 < u \leq N_{user}; 0 < p \leq N_{page}$$



Hình 2.6: Ví dụ về đồ thị trung gian

Từ đồ thị trên, tính được số trang web ghé thăm chung của 2 người dùng u_p và u_q là n_{page}

$$n_{page}(u_p, u_q) = \sum_{k=1}^{N_{page}} P[u_p, k].P[u_q, k]$$

Tương tự, Gọi N_{page}, N_{cate} và N_{topic} là tổng số trang, tổng số chuyên mục và tổng số chủ đề trong trang web tương ứng. Và n_{page} (u_p, u_q), n_{cate} (u_p, u_q), n_{topic} (u_p, u_q) tương ứng là số lượt truy cập cùng trang, đánh số lượt truy cập cùng chuyên mục và số lượt truy cập cùng chủ đề của cả người dùng u_p và u_q, để tính được n_{cate} và n_{topic} ta cần tạo bảng ánh xạ giữa trang web và chuyên mục và chủ đề của trang web đó.

Bảng 2.5: Ánh xạ giữa trang web và chuyên mục, chủ đề

Trang web	Chuyên mục	Chủ đề
/tin-tuc/1	Tin tức	Chủ đề 2
/tin-tuc/2	Tin tức	Chủ đề 2
/thong-bao/1	Thông báo	Chủ đề 1
/giao-duc/1	Giáo dục	Chủ đề 3
/kinh-te/1	Kinh tế	Chủ đề 3
…	…	…
/thoi-su/1	Thời sự	Chủ đề 1

Xác định đồ thị trung gian theo chuyên mục, chủ đề

$$C[u, c] = \begin{cases} 1, & nếu\ u\ truy\ cập\ c \\ 0, & ngược\ lại \end{cases}$$

$$với\ \ 0 < u \leq N_{user}; 0 < p \leq N_{cate}$$

$$T[u, t] = \begin{cases} 1, & nếu\ u\ truy\ cập\ t \\ 0, & ngược\ lại \end{cases}$$

$$với\ \ 0 < u \leq N_{user}; 0 < p \leq N_{topic}$$

Số chuyên mục, chủ đề chung giữa hai người dùng được tính như sau:

$$n_{cate}(u_p, u_q) = \sum_{k=1}^{N_{cate}} C[u_p, k].C[u_q, k]$$

$$n_{topic}(u_p, u_q) = \sum_{k=1}^{N_{topic}} T[u_p, k].T[u_q, k]$$

Mối quan tâm của người dùng có thể được định nghĩa là một tập hợp các trang, một tập hợp các chuyên mục, hoặc một tập hợp các chủ để mà người dùng đã truy cập. Đối với mỗi lần truy cập trang, chuyên mục hay các chủ đề, ta gia sử số lần truy cập t lớn hơn một ngưỡng T. Nếu t nhỏ hơn hoặc bằng T thì có nghĩa là người dùng không có mối quan tâm đến trang web, chuyên mục hay chủ đề này.

Để đơn giản, giả sử khoảng thời gian người dùng truy cập một chuyên mục hoặc một chủ đề trên trang web là khoảng thời gian người dùng duyệt trang này. Dựa trên những giả định này, việc truy cập một trang web, lượt truy cập của một chuyên mục và lượt truy cập của một chủ đề được sử dụng như ba chỉ số chính để đo lường sở thích của người dùng. Các chỉ số được xác định như sau.

Lượt truy cập theo trang web: Lượt truy cập trang cho biết hoạt động duyệt trang web p_i qua URL của người dùng trong một khoảng thời gian nhất định t > T.

Bằng cách xác định độ tương tự giống như số trang web giữa hai người dùng cùng ghé thăm, có thể sử dụng chuyên mục, chủ đề mà hai người dùng cùng ghé thăm để xác định điểm chung giữa hai người dùng. Cách xác định này cũng xây dựng được một đồ thị tương tự giữa người dùng trong hệ thống website giống như ma trận xác định trên số trang web cùng ghé thăm nhưng kết quả phân cụm có thể khác nhau.

Lượt truy cập theo chuyên mục: Lượt truy cập chuyên mục cho biết hoạt động duyệt một trang web p_i thuộc về một chuyên mục ctg_j bởi người dùng trong một khoảng thời gian nhất định t > T.

Có những trường hợp một bài đăng trong một trang web phải thuộc nhiều chuyên mục nhưng quản trị viên được chỉ định một hoặc chưa gán chuyên mục, thì

chuyên mục này có thể không phù hợp để mô tả mối quan tâm của người dùng. Chúng ta có thể dựa vào chủ đề của một trang web. Chủ đề này được tính toán bằng cách sử dụng các kỹ thuật mô hình chủ đề trên tiêu đề của các trang web trong trang web. Mô hình chủ đề là một kỹ thuật để trích xuất các chủ đề ẩn từ khối lượng lớn văn bản, sử dụng LDA. Ở đây, dựa vào các chủ đề này để xác định mối quan tâm của người dùng, thông qua chủ đề được xác định như sau.

Lượt truy cập theo chủ đề: Lượt truy cập chủ đề là một chủ đề cho biết hoạt động duyệt một trang web p_i thuộc về một chủ đề tpc_j bởi người dùng trong một khoảng thời gian nhất định t > T. Trong đó một tpc_j chủ đề được tính bằng cách áp dụng mô hình LDA được xây dựng từ một tập hợp các tiêu đề trích xuất của các trang web của trang web trên tiêu đề của trang duyệt web.

Từ mỗi chỉ số được xác định ở trên có thể xây dựng các đồ thị tương tự, trong đó mỗi nút là một người dùng và các cạnh được đánh giá dựa trên sự tương đồng giữa hai lần truy cập trang web của người dùng, lượt truy cập chuyên mục hoặc lượt truy cập chủ đề. Sau đó, xác định ba hàm để đo lường mức độ tương tự giữa hai người dùng về mối quan tâm của người dùng như sau.

Độ tương tự của lượt truy cập trang: Độ tương tự của lượt truy cập trang giữa hai người dùng và u_q phụ thuộc vào n_{page} (u_p, u_q) và được tính như sau:

$$sim_{page\_visit}(u_p, u_q) = \begin{cases} \dfrac{n_{page}}{N_{page}} & \dfrac{n_{page}}{N_{page}} \geq \alpha_{page} \\ 0 & \text{ngược lại} \end{cases}$$

trong đó α_{page} là giá trị ngưỡng và $0 \leq sim_{page\_visit} \leq 1$.

Tương tự, độ tương tự truy cập chuyên mục và độ tương tự truy cập chủ đề được tính như sau:

$$sim_{cate\_visit}(u_p, u_q) = \begin{cases} \dfrac{n_{cate}}{N_{cate}} & \dfrac{n_{cate}}{N_{cate}} \geq \alpha_{cate} \\ 0 & \text{ngược lại} \end{cases}$$

$$sim_{topic\_visit}(u_p, u_q) = \begin{cases} \dfrac{n_{topic}}{N_{topic}} & \dfrac{n_{topic}}{N_{topic}} \geq \alpha_{topic} \\ 0 & \text{ngược lại} \end{cases}$$

trong đó α_{cate}, α_{topic} là giá trị ngưỡng và $0 \leq sim_{cate\_visit}, sim_{topic\_visit} \leq 1$.

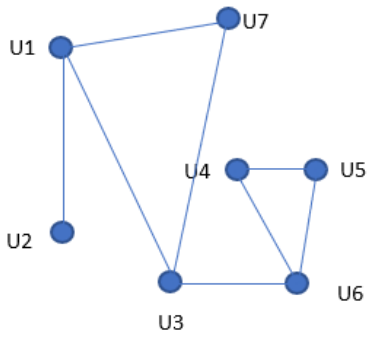
Giá trị của các độ đo tương tự ở trên nằm trong khoảng từ 0 đến 1 và khoảng cách cao cho thấy độ tương tự cao giữa mối quan tâm của hai người dùng. Nếu giá trị là 0, không có sự tương đồng và người dùng không quan tâm đến trang web, chuyên mục hoặc chủ đề. Sử dụng ba loại mức độ tương tự giữa hai người dùng, ta có thể xây dựng ba loại đồ thị tương tự: đồ thị trang web, đồ thị chuyên mục và đồ thị chủ đề.

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 này là đồ thị vô hướng có số đỉnh chính là số người dùng đã truy cập web dựa trên log đã thu thập được. Cạnh của đồ thị có trọng số thể hiện độ tương tự của người dùng. Trọng số này càng lớn thì người dùng có nhiều điểm tương tự. Dựa và các độ tương tự đã được tính toán ở trên, ta xác định được ba loại 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 2.7: 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. Nó sẽ cho kết quả những người dùng có cùng mối quan tâm sẽ được đặt trong cùng một cụm, những người có mối quan tâm, sở thích khác nhau sẽ được đặt trong các cụm khác nhau.

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.

Bắt đầu với đồ thị tương tự của tất cả người dùng. Lần đầu tiên chia nhóm đồ thị tương tự, ta sẽ có được cụm cao cấp nhất, ví dụ cụm 1 và cụm 2. Thuật toán phân cụm được sử dụng là phân cụm đồ thị phân cấp bằng cách lấy mẫu cặp nút[10]. Đây là một thuật toán phân cụm liên kết cho các đồ thị dựa trên khoảng cách có thể rút ngắn giữa các cụm và có thể cung cấp một hệ thống phân cấp đầy đủ của đồ thị. Thuật toán nắm bắt được cấu trúc với quy mô khác nhau của đồ thị thực, tham số tự do, nhanh chóng và hiệu quả.

Sử dụng thuật toán phân cụm này, ta tiếp tục chia các cụm 1 và cụm 2 thành các cụm con nhỏ hơn. Ví dụ: cụm 3, cụm 4 là 2 cụm con của cụm 1. Quá trình phân cụm này được áp dụng lặp lại cho các cụm mới được sinh ra. Nó dừng lại khi giá trị *modularity* của các cụm đạt ngưỡng tối thiểu. Giá trị *modularity* là một số liệu đo mật độ của các cạnh trong cụm đến các cạnh bên ngoài cụm [12]. Kết quả là một hệ thống cây phân cấp các cụm mối quan tâm của người dùng.

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.

Ví dụ: Nhóm người dùng thường xuyên đọc các trang web thuộc các chuyên mục: bóng đá, quần vợt, bóng chuyền,… có thể thấy được đây là nhóm người dùng quan tâm đến lĩnh vực thể thao.

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.

Ví dụ: Nhóm người dùng thường xuyên đọc các trang web có liên quan đến các từ khóa: học bổng, tuyển sinh, trung học phổ thông,… có thể là nhóm người dùng quan tâm đến lĩnh vực giáo dục.

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.

Ngoài ra, do kết quả phân cụm là cây phân cấp của các cụm mối quan tâm của người dùng, có thể phân tích các cấp độ cụm khác nhau theo mối quan tâm của người

dùng. Điều này giúp tiết lộ những hiểu biết chính về các nhóm người dùng cũng như mối quan tâm của họ. Sau đó, kết hợp tất cả các kết quả phân tích sau khi thực hiện trên ba loại đồ thị tương tự được đề xuất, chúng ta có thể hiểu thêm về mối quan tâm của người dùng trên trang web.

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

2.5. Kết luận chương

Chương 2 đã trình bày về cách xây dựng giải pháp kỹ thuật thu thập 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 - THỰC NGHIỆM VÀ KẾT QUẢ

Xây dựng bộ dữ liệu từ dữ liệu thực tế, sử dụng kỹ thuật học không giám sát đã đề xuất ở chương 2 để đưa ra kết quả phân cụm người dùng. Dựa vào kết quả thu được đưa ra các đánh giá cho các cụm người dùng.

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

– Tải package Linux Binaries 64bit từ trang chủ của NodeJS

```
# wget https://nodejs.org/dist/v8.9.3/node-v8.9.3-linux-x64.tar.xz
```

– Giải nén nội dung bên trong vào /usr/local

```
# tar --strip-components 1 -xJvf node-v8.9.3-linux-x64.tar.xz -C /usr/local
```

– Kiểm tra lại phiên bản NodeJS

```
# node --version

v8.9.3
```

Cài đặt MongoDB

– Import "MongoDB public GPG Key" sử dụng command apt-key

```
# sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 9DA31620334BD75D9DCB49F368818C72E52529D4

# echo "deb [ arch=amd64,arm64 ] https://repo.mongodb.org/apt/ubuntu xenial/mongodb-org/4.0 multiverse" | sudo tee /etc/apt/sources.list.d/mongodb-org-4.0.list

# sudo apt-get update
```

– Cài đặt

```
# sudo apt-get install -y mongodb-org
```

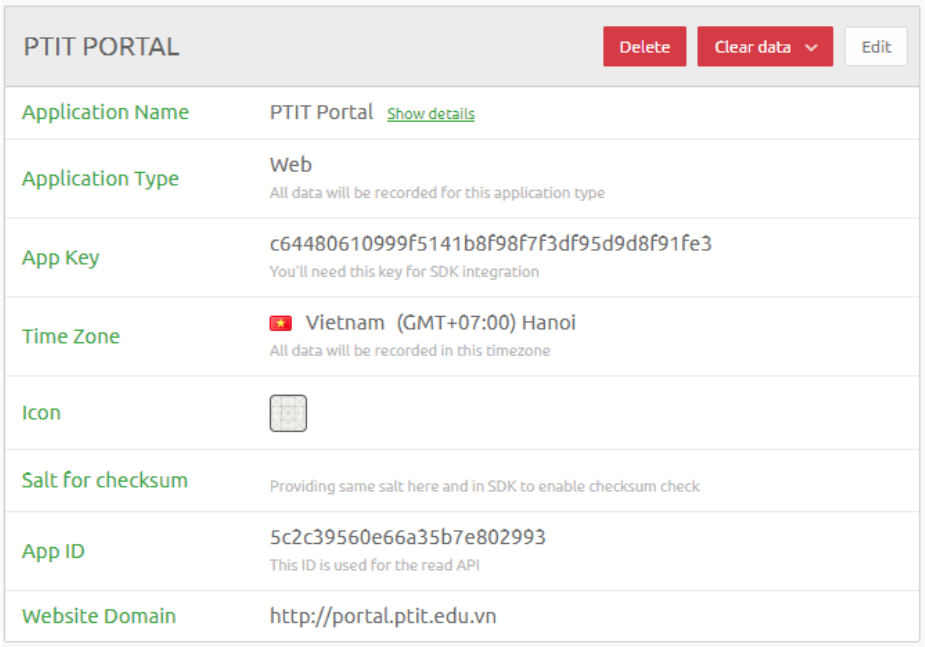
– Khởi động và Kiểm tra lại phiên bản MongoDB

```
# sudo service mongod start

# mongodb --vesion

db version v.4.0.8
```

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.

```
# sudo su -

# wget -qO- http://c.ly/install | bash
```

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.



Hình 3.1: Thông tin chi tiết ứng dụng cần thu thập log

Cuối cùng, cần sinh mã nhũng 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 -->
```

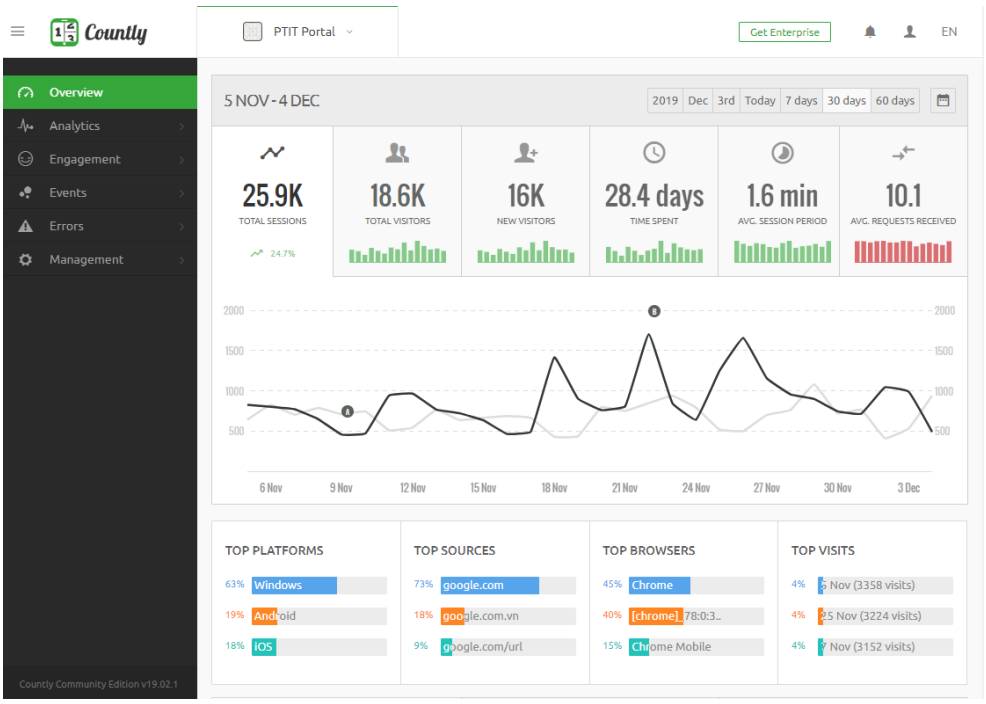
Hình 3.2: Mã nhúng tích hợp dành cho website cần thu thập

Khi nhúng mã theo dõi lên website, có hai lựa chọn là mã nhúng đồng bộ và mã nhúng bất đồng bộ. Ta nên sử dụng mã nhúng bất đồng bộ vì có lợi thế cải thiện được tốc độ tải trang, các dữ liệu về tương tác của người dùng vẫn được thu thập và đẩy vào hàng đợi ngay cả khi mã nhúng chưa được tải xong. Sử dụng mã nhúng bất đồng bộ cũng không gây ảnh hưởng đến website được theo dõi, ngay cả khi máy chủ thu thập log gặp sự cố, website vẫn có thể hoạt động bình thường.

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.



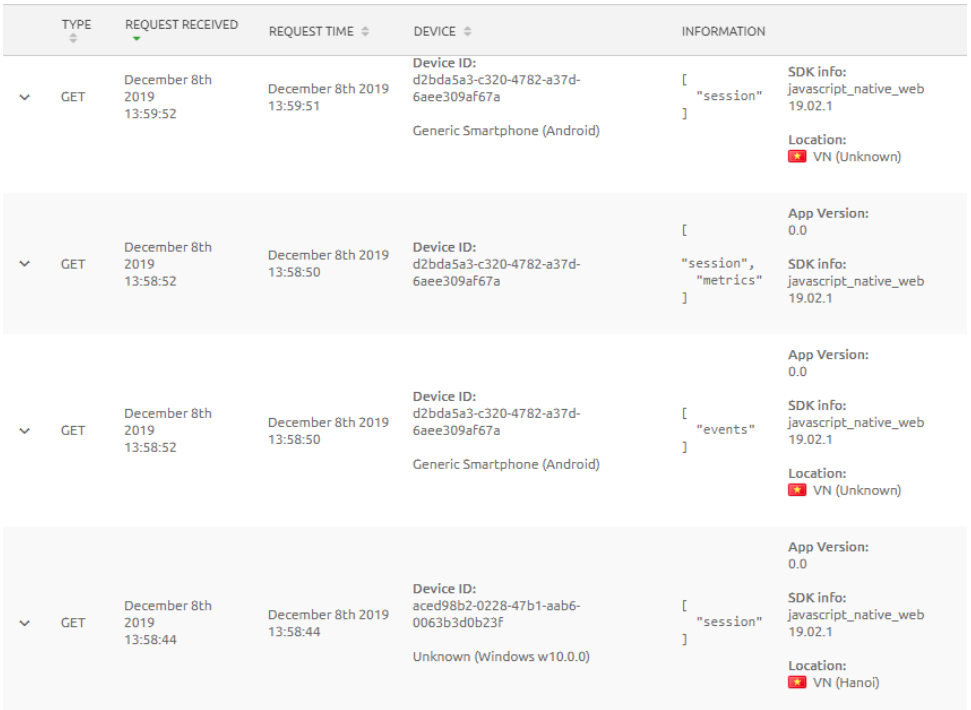
Hình 3.3: Giao diện thống kê truy cập trong khoảng một thời gian

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, … Các trang web con thường là các bài đăng có nội dung chủ yếu là văn bản, ví dụ như các thông báo từ các phòng ban trong Học viện tới sinh viên, các thông tin về các hoạt động câu lạc bộ, hoạt động của sinh viên trong và ngoài học viện. Ngoài ra, theo từng giai đoạn trong năm học, có thời điểm tập trung nhiều vào các trang web có nội dung về tuyển sinh (thời gian tuyển sinh khóa mới theo kế hoạch của Bộ Giáo dục và Đào tạo), hoặc nội dung về kế hoạch thi, lịch thi, điểm thi của sinh viên (giai đoạn cuối học kỳ),...

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 địa chỉ trang web, thời gian và nội dung chi tiết của từng trang web được ghi lại trong bộ dữ liệu. Tất cả dữ liệu lượt truy cập của người dùng tương tự được gán cho một ID thiết bị duy nhất.



	TYPE	REQUEST RECEIVED	REQUEST TIME	DEVICE	INFORMATION	
⌄	GET	December 8th 2019 13:59:52	December 8th 2019 13:59:51	Device ID: d2bda5a3-c320-4782-a37d-6aee309af67a Generic Smartphone (Android)	["session"]	SDK info: javascript_native_web 19.02.1 Location: VN (Unknown)
⌄	GET	December 8th 2019 13:58:52	December 8th 2019 13:58:50	Device ID: d2bda5a3-c320-4782-a37d-6aee309af67a	["session", "metrics"]	App Version: 0.0 SDK info: javascript_native_web 19.02.1
⌄	GET	December 8th 2019 13:58:52	December 8th 2019 13:58:50	Device ID: d2bda5a3-c320-4782-a37d-6aee309af67a Generic Smartphone (Android)	["events"]	App Version: 0.0 SDK info: javascript_native_web 19.02.1 Location: VN (Unknown)
⌄	GET	December 8th 2019 13:58:44	December 8th 2019 13:58:44	Device ID: aced98b2-0228-47b1-aab6-0063b3d0b23f Unknown (Windows w10.0.0)	["session"]	App Version: 0.0 SDK info: javascript_native_web 19.02.1 Location: VN (Hanoi)

Hình 3.4: Chi tiết về dữ liệu thu thập được từ người dùng

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. Dữ liệu sau khi được tiền xử lý được lưu trữ trong CSDL với MongoDB. 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 [11]. Thông thường, nó không quá 180 giây [9].

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 dưới đây

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) [2] 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.

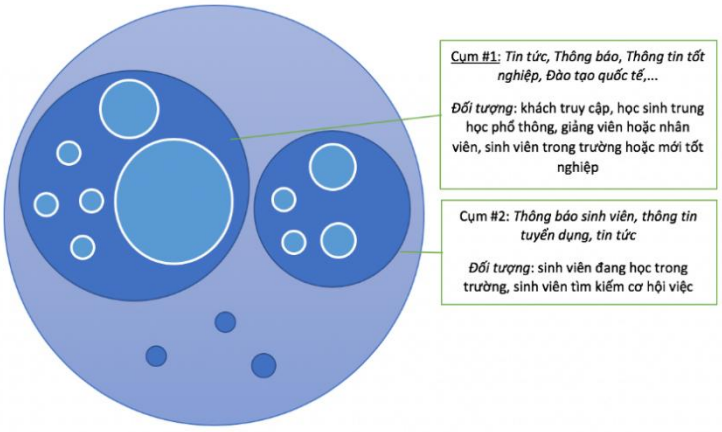
max.diameter là khoảng cách lớn nhất trong nội bộ cụm (giống như đường kính).

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 3.5: 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.

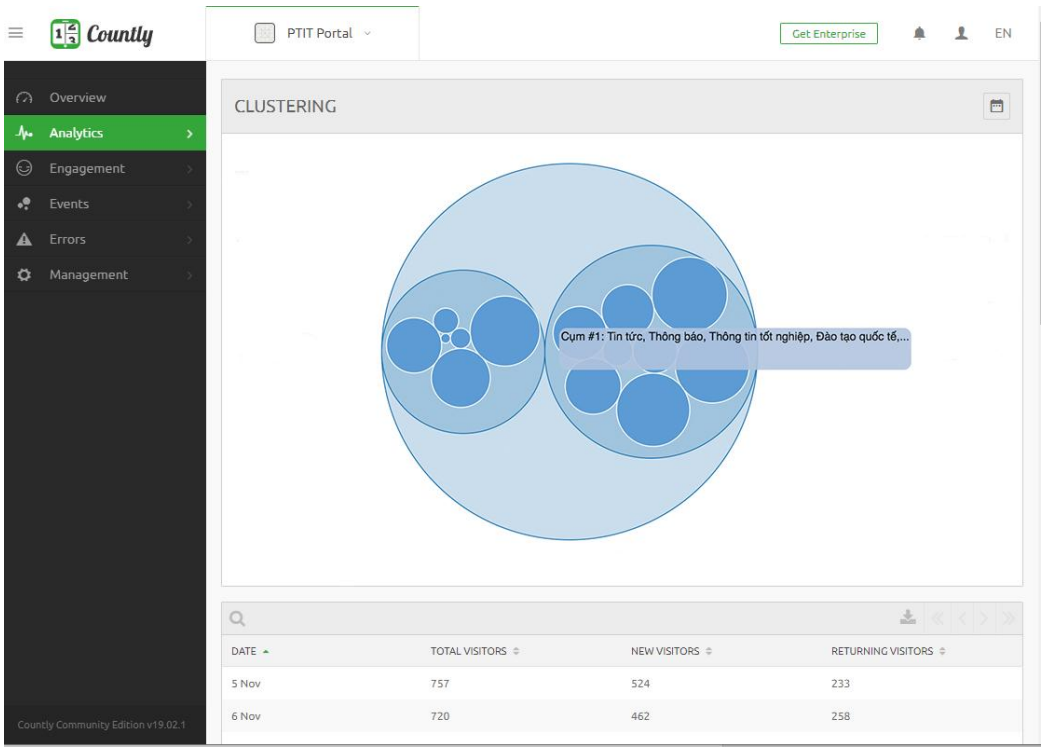
Đồ thị theo trang web. Áp dụng thuật toán phân cụm phân cấp vào đồ thị theo trang web, người dùng được phân thành 7 cụm. Sau đó, đối với mỗi trang web, ánh xạ tới chuyên mục và chủ đề tương ứng. Bảng 3.6 mô tả ba cụm trên cùng trong kết quả phân cụm sau khi gán tên chuyên mục. Từ kết quả, chỉ biết rằng 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 3.6: 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.

Sau khi quản trị viên thực hiện thao tác phân cụm người dùng. Một tiến trình sẽ thực hiện ngầm, sau khi thực hiện xong thuật toán phân cụm, dữ liệu sẽ được lưu trữ ở bộ nhớ đệm, sau đó sử dụng các công cụ biểu diễn biểu đồ để biểu diễn các cụm dữ liệu.

Từ biểu đồ, quản trị viên có thể xem xét về sở thích, các mối quan tâm, các nhóm người dùng trên website để có thể tiến hành những thay đổi cần thiết cho nội dung phù hợp với người dùng hơn.

3.3. 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 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.

Do thời gian thực hiện luận văn không nhiều nên tác giả chưa có điều kiện nghiên cứu thêm nhiều phương pháp. Trong tương lai, nếu có điều kiện, tác giải sẽ tập trung nghiên cứu để 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.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Hoàng Văn Dũng, (2007), *Khai phá dữ liệu web bằng kỹ thuật phân cụm*, Hà Nội, pp. 31-33.

[2] Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta, (2008), *ClValid: An R Package for Cluster Validation*, Journal of Statistical Software 25 (4), pp. 1–22.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, (2003), *Latent Dirichlet allocation*, J. Mach. Learn. Res, pp. 996-999.

[4] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, Ben Y. Zhao, (2016) *Unsupervised Clickstream Clustering for User Behavior Analysis*. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 225-236.

[5] Jaideep Srivastava, Robert Cooley y, Mukund Deshpande, Pang-Ning Tan, (2000), *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. In SIGKDD Explorations, Volume 1, pp. 2-4.

[6] Justin Cutroni, (2010), *Google Analytics: Understanding Visitor Behavior*, Chapter 3, pp. 13-19.

[7] L.K. Joshila Grace, V. Maheswari, and Dhinaharan Nagamalai, (2011), *Analysis of Web Logs And Web User In Web Mining*, In International Journal of Network Security & Its Applications, Volume 3, pp. 99-101.

[8] Peter Zadrozny, Raghu Kodali, (2013), *Big Data Analytics Using Splunk*. pp. 31-33.

[9] Q. Su and L. Chen, (2015) *A method for discovering clusters of e-commerce interest patterns using click-stream data*, Electron. Commer. Res. Appl, pp. 6-7.

[10] Thomas Bonald, Bertrand Charpentier, Alexis Galland, (2018), *Alexandre Hollocou. Hierarchical Graph Clustering using Node Pair Sampling*, In MLG 2018 - 14th International Workshop on Mining and Learning with Graphs, London, United Kingdom, pp. 1-3.

[11] T. Haile, (2014) *What you think you know about the web is wrong*, Time. com, March, vol. 9.

[12] U. Brandes et al, (2008) *On modularity clustering*, IEEE Trans. Knowl. Data Eng, pp. 3-6.