

ỨNG DỤNG LỌC CỘNG TÁC TRONG WEBSITE BÁN HÀNG THỜI TRANG USING COLLABORATIVE FILTERING IN FASHION ECOMMERCE WEBSITE

Trần Lê Công Hậu¹, Nguyễn Thị Hoài Linh²

¹Trường Đại học Kinh tế - Tài chính, Hồ Chí Minh, hautlc19@uef.edu.vn,

²Trường Đại học Kinh tế - Tài chính, Hồ Chí Minh, linhnth@uef.edu.vn

Tóm tắt: Lọc cộng tác là một phương pháp sử dụng thông tin từ nhóm người dùng để tạo ra các gợi ý, đề xuất hoặc phân loại cho một người dùng cụ thể. Thông qua việc thu thập và phân tích dữ liệu từ các người dùng khác, lọc cộng tác đánh giá và dự đoán sở thích, hành vi hoặc sự tương tác của một người dùng mà không yêu cầu thông tin cá nhân chi tiết. Bài báo nghiên cứu về lọc cộng tác trong hệ thống gợi ý dựa trên lịch sử người dùng và cách thức tích hợp nó vào website bán hàng thời trang nhằm nâng cao trải nghiệm của khách hàng, góp phần tăng doanh thu cho doanh nghiệp.

Từ khóa: Lọc cộng tác, hệ thống gợi ý, website bán hàng thời trang, singular value decomposition.

Abstract: Collaborative filtering is a method that utilizes information from a group of users to generate recommendations, suggestions, or classifications for a specific user. By collecting and analyzing data from other users, collaborative filtering evaluates and predicts the preferences, behaviours, or interactions of a user without requiring detailed personal information. The research paper focuses on collaborative filtering in recommendation systems based on user history and how to integrate it into a fashion e-commerce website to enhance the customer experience and contribute to increased business revenue.

Keywords: Collaborative filtering, recommendation system, fashion ecommerce website, singular value decomposition.

1. Giới thiệu

Đề tài "Ứng dụng lọc cộng tác trong website bán hàng thời trang" là một đề tài nghiên cứu trong lĩnh vực công nghệ thông tin, tập trung vào việc áp dụng công nghệ lọc cộng tác vào các website bán hàng thời trang. Đề tài nhằm nghiên cứu và phân tích các công nghệ lọc cộng tác, xây dựng một hệ thống lọc cộng tác cho website bán hàng thời trang dựa trên các công nghệ mới nhất và đánh giá hiệu quả của hệ thống lọc cộng tác được xây dựng trong việc tăng cường trải nghiệm người dùng và độ chính xác trong việc gợi ý sản phẩm.

2. Hệ thống gợi ý

2.1. Khái niệm

Hệ thống gợi ý là một hệ thống thông tin tự động giúp đề xuất, gợi ý hoặc đưa ra các thông tin, sản phẩm hoặc dịch vụ phù hợp với sở thích, hành vi mua sắm, thông tin cá nhân của người dùng. Nó thường được sử

dụng trong nhiều lĩnh vực khác nhau như mua sắm trực tuyến, xem phim, nghe nhạc, đọc sách và tài liệu, và các ứng dụng di động.

2.2. Ví dụ ma trận tiện ích (utility matrix)

Trong hệ thống này, hai thực thể cơ bản luôn xuất hiện trong bất kỳ hệ thống gợi ý đó là người dùng và chỉ mục. Thực thể người dùng là những người sử dụng hệ thống, để lại các lịch sử sử dụng như đưa ra các ý kiến của cá nhân về các chỉ mục khác nhau và nhận đề xuất về các chỉ mục mới từ hệ thống. Các chỉ mục đại diện cho các mặt hàng, sản phẩm, bài báo,... có trong hệ thống, cần được thông tin đến người dùng.

Như đề cập ở trên thì hai thực thể chính trong hệ thống gợi ý là người dùng (user) và chỉ mục (item). Mỗi người dùng u có mức độ quan tâm chỉ mục i khác nhau, giá trị được đánh giá là r_{ui} , đây là giá trị liên kết giữa người dùng u và chỉ mục i . Mức độ quan tâm này, nếu đã biết trước, được gán cho một giá

Bảng 1. Ví dụ về ma trận tiện ích với hệ thống gợi ý thời trang

	user1	user2	user3	user4	user5	user6
Áo dài	5	5	0	0	1	?
Váy	5	?	?	0	?	?
Vòng tay	?	4	1	?	?	1
Máy tính	1	1	4	4	4	?
Điện thoại	1	0	5	?	?	?

trị ứng với mỗi cặp người dùng-chỉ mục. Giả sử rằng mức độ quan tâm được đo bằng giá trị người dùng đánh giá cho chỉ mục, ta tạm gọi giá trị này là điểm đánh giá (rating). Tập hợp tất cả các ratings, bao gồm cả những giá trị chưa biết cần được dự đoán, tạo nên một ma trận gọi là utility matrix.

Ví dụ trong Bảng 1, có 6 người dùng và có 5 chỉ mục. Các ô có giá trị thể hiện việc một người dùng đã đánh giá một chỉ mục với ratings từ 0 (không thích) đến 5 (rất thích). Các ô có dấu ‘?’ tương ứng với các ô chưa có dữ liệu. Công việc của một hệ thống gợi ý là dự đoán giá trị tại các ô này, từ đó đưa ra gợi ý cho người dùng. Vì vậy, đôi khi cũng được coi là bài toán Hoàn thiện ma trận (Matrix Completion).

2.3. Phân loại

Lọc dựa trên nội dung: Là quá trình sử dụng các thuật toán và phương pháp để đánh giá và phân loại nội dung dựa trên các yếu tố nội dung như từ khóa, thể loại, thông tin mô tả và các thuộc tính khác của tài liệu. Mục đích của việc lọc này là tìm ra các tài liệu có nội dung liên quan và phù hợp nhất để đề xuất cho người dùng.

Lọc cộng tác: Là quá trình sử dụng thông tin phản hồi và đánh giá từ người dùng để tạo ra các đề xuất và gợi ý cá nhân hóa. Thay vì chỉ dựa trên nội dung, lọc cộng tác sử dụng thông tin về hành vi, sở thích, và phản hồi của người dùng để đề xuất nội dung phù hợp.

Lọc dựa trên cơ sở tri thức: Là quá trình sử dụng kiến thức, thông tin và quy tắc được thu thập và tổ chức trong một cơ sở tri thức để đánh giá và phân loại nội dung, sau đó tạo ra các đề xuất và gợi ý cho người dùng dựa trên kiến thức này.

Lọc dựa trên kết hợp các công nghệ lọc: Là một phương pháp kết hợp sử dụng cả lọc dựa trên nội dung và lọc cộng tác để cung cấp các đề xuất và gợi ý cho người dùng.

3. Lọc cộng tác trong website bán hàng

Trong phương pháp lọc cộng tác, có nhiều tiếp cận khác nhau có thể được áp dụng như đề cập ở mục 2.3. Tuy nhiên, trong trường hợp này, chúng tôi chọn sử dụng phân tích dựa trên lịch sử người dùng.

3.1. Phân tích

Lọc cộng tác dựa trên của lịch sử người dùng là một phương pháp trong hệ thống gợi ý để tìm ra các mục tiêu hoặc nội dung phù hợp với người dùng dựa trên hành vi và sở thích của người dùng trong quá khứ. Ý tưởng chính là nếu hai người dùng có lịch sử hoạt động tương tự, thì có khả năng cao rằng họ sẽ có sự quan tâm chung đến nội dung hoặc mục tiêu tương tự.

Bảng.2. Bảng dữ liệu website đã được chuẩn hóa

User	Item	Bought
UserId1	I1	1.0
UserId2	I2	1.0
UserId3	I2	1.0

Dữ liệu ở bảng 2. đã được chuẩn hóa với UserId1, UserId2, UserId3 là các mã người dùng, I1, I2, I3 là các chỉ mục (mặt hàng). Bought với giá trị 0/1 đánh dấu việc user đã mua item hay chưa dựa theo lịch sử của người dùng.

3.2. Cơ bản về phân tích giá trị đơn (Singular Value Decomposition)

SVD, viết tắt của Singular Value Decomposition là một phương pháp phân tích ma trận được sử dụng rộng rãi trong lĩnh vực xử lý dữ liệu và đại số tuyến tính. SVD cho phép một ma trận được phân tích thành các thành phần chính, giúp tái tạo lại ma trận ban đầu và tìm ra các đặc trưng quan trọng của dữ liệu.

Phân tích SVD giải phương trình:

$$A = U\Sigma V^T$$

Trong đó A là một ma trận có kích thước $m \times n$, U là ma trận trực giao kích thước $m \times m$, Σ là ma trận đường chéo khối $m \times n$ và V là ma trận trực giao kích thước $n \times n$.

Các ma trận U và V trực giao trong SVD có vai trò quan trọng trong việc tách riêng

HỘI THẢO NGHIÊN CỨU KHOA HỌC SINH VIÊN KHOA CNTT LẦN 1 NĂM 2023
ĐỔI MỚI SÁNG TẠO VÀ HỘI NHẬP QUỐC TẾ TRONG THỜI ĐẠI 4.0

các thành phần chính của ma trận A. Ma trận U chứa các vector cột riêng của ma trận A, gọi là vector-sing, và ma trận V chứa các vector cột riêng của ma trận chuyển vị A, gọi là vector-sing phụ. Các giá trị trên đường chéo của ma trận Σ , gọi là giá trị singular, đại diện cho mức độ quan trọng của từng thành phần trong phân tích.

Tiếp theo, dưới đây là ví dụ sử dụng SVD trong hệ thống lọc cộng tác gợi ý cho userId5(1,?,?,?):

Bước 1: Xây dựng ma trận dữ liệu

	I1	I2	I3	I4
userId1	1	0	1	0
userId2	0	1	1	1
userId3	0	1	0	0
userId4	0	1	0	1
userId5	1	?	?	?

Bảng 3. Bảng ma trận dữ liệu

Bước 2: Chuẩn hóa ma trận dữ liệu, giá trị của mỗi ô trong hàng sẽ bằng tổng giá trị của hàng đó chia cho số lượng ô có giá trị.

	I1	I2	I3	I4
userId1	0.5	-0.5	0.5	-0.5
userId2	-0.75	0.25	0.25	0.25
userId3	-0.25	0.75	-0.25	-0.25
userId4	-0.5	0.5	-0.5	0.5
userId5	1	0	0	0

Bảng 4. Bảng ma trận chuẩn hóa dữ liệu

Bước 3: Áp dụng thuật toán SVD

0.5	-0.5	0.5	-0.5
-0.75	0.25	0.25	0.25
-0.25	0.75	-0.25	-0.25
-0.5	0.5	-0.5	0.5
1	0	0	0

Bảng 5. Khởi tạo ma trận A

Sử dụng thuật toán SVD phân rã ma trận A, chúng ta có các ma trận U, Σ , V^T như sau:

-0.53	-0.33	0.33	0.045	-0.71
0.398	-0.47	0.053	0.79	0
0.314	0.38	0.87	0.007	0
0.53	0.33	-0.334	-0.049	-0.71
-0.43	0.65	-0.1315	0.61	0

Bảng 6. Bảng các giá trị của ma trận U

1.76	0.94	0.65	0.33
------	------	------	------

Bảng 7. Bảng các giá trị riêng Σ

-0.76	0.49	-0.29	0.31
0.61	0.53	-0.58	0.13
-0.09	0.51	0.2	-0.83

0.2	0.46	0.74	0.44
-----	------	------	------

Bảng 8. Bảng các giá trị của V^T

Bước 4: Giảm chiều dữ liệu, với k=2

-0.53	-0.33
0.398	-0.47
0.314	0.38
0.53	0.33
-0.43	0.65

Bảng 9. Bảng các giá trị của ma trận U'

1.76	0.94
------	------

Bảng 10. Bảng các giá trị riêng Σ'

-0.76	0.49
0.61	0.53
-0.09	0.51
0.2	0.46

Bảng 11. Bảng các giá trị của ma trận V'^T

Bước 5: Tính độ tương đồng

$$\text{cosine_similarity}(\mathbf{u}_1, \mathbf{u}_2) = \cos(\mathbf{u}_1, \mathbf{u}_2) = \frac{\mathbf{u}_1^T \mathbf{u}_2}{\|\mathbf{u}_1\|_2 \cdot \|\mathbf{u}_2\|_2}$$

$$(\text{user5}, \text{user1}) = 0.0275391$$

$$(\text{user5}, \text{user2}) = -0.993023$$

$$(\text{user5}, \text{user3}) = 0.291476$$

$$(\text{user5}, \text{user4}) = -0.02$$

Với k=2, suy ra, chúng ta sẽ gợi ý I2, I3 cho userId5.

Sử dụng SVD, ta có thể áp dụng nén dữ liệu, giảm chiều dữ liệu, xác định các đặc trưng quan trọng và thực hiện các phép tính tương tự trong không gian tạo bởi các thành phần chính của ma trận A.

Tuy SVD có ứng dụng rộng rãi trong nhiều lĩnh vực như xử lý ảnh, nhận dạng mẫu, khuyến nghị, nhưng phép tính SVD thường có độ phức tạp cao và yêu cầu tài nguyên tính toán lớn. Do đó, trong thực tế, các phương pháp ước lượng SVD và biểu diễn SVD bằng cách giảm kích thước ma trận thường được sử dụng để tối ưu hiệu suất tính toán.

3.3. Lý do lựa chọn thuật toán phân tích giá trị đơn

Giảm chiều dữ liệu: SVD cho phép chúng ta giảm chiều dữ liệu mà vẫn giữ lại thông tin quan trọng. Khi áp dụng SVD lên ma trận đánh giá của người dùng và sản phẩm, chúng ta có thể tạo ra một biểu diễn gọn hơn của dữ

liệu. Điều này giúp giảm độ phức tạp tính toán và tiết kiệm tài nguyên.

Khám phá sở thích ẩn: SVD có khả năng khám phá những sở thích ẩn trong dữ liệu. Bằng cách phân tích các giá trị suy biến, chúng ta có thể tìm ra các mô hình ẩn đằng sau sự tương tác giữa người dùng và sản phẩm. Điều này giúp xác định những mối quan hệ và sự tương đồng giữa các người dùng và các sản phẩm, từ đó đưa ra gợi ý phù hợp.

Xây dựng hệ thống gợi ý: SVD là một công cụ quan trọng trong xây dựng hệ thống gợi ý. Bằng cách sử dụng SVD, chúng ta có thể xây dựng các mô hình gợi ý dựa trên thông tin về đánh giá của người dùng và sản phẩm. Mô hình này có thể dự đoán các đánh giá chưa được biết đến và gợi ý các sản phẩm tương tự dựa trên sở thích của người dùng.

Hỗ trợ lọc cộng tác: SVD được sử dụng trong lọc cộng tác, một phương pháp quan trọng trong hệ thống gợi ý. Bằng cách áp dụng SVD lên ma trận đánh giá, chúng ta có thể tìm ra sự tương đồng giữa các người dùng hoặc giữa các sản phẩm. Điều này cho phép chúng ta đưa ra những gợi ý dựa trên hành vi và sở thích của người dùng tương tự hoặc các sản phẩm có thuộc tính tương đồng.

4. Triển khai

4.1. Quy trình xây dựng hệ thống gợi ý

Cài đặt môi trường ảo Anaconda để có thể sử dụng được các gói của Python, sử dụng được các công cụ phân tích dữ liệu.

Chuẩn bị Jupyter Notebook: Mở Jupyter Notebook và tạo một notebook mới để làm việc với dữ liệu. Cài đặt các thư viện và gói phần mềm cần thiết như pandas, numpy, scikit-learn để thực hiện các công việc làm sạch và huấn luyện dữ liệu.

Import dữ liệu: Sử dụng pandas để import dữ liệu từ các nguồn như file CSV, Excel, hoặc cơ sở dữ liệu SQL. Đọc dữ liệu vào DataFrame để tiện cho việc làm việc và xử lý. Dữ liệu đầu vào là dữ liệu về các khách hàng, lịch sử giao dịch, danh sách các sản phẩm bao gồm các file như sau: customers.csv (197MB), transactions.csv (3.4GB), products.csv (30MB) được tham khảo từ website kaggle [11].

Khám phá dữ liệu: Tiến hành khám phá dữ liệu để hiểu về cấu trúc, thông tin và đặc điểm của khách hàng. Sử dụng các phương pháp như head(), shape() để xem thông tin về dữ liệu, loại bỏ các giá trị null, undefined.

Xử lý dữ liệu: Những giá trị bị thiếu, không hợp lệ sẽ bị xóa đi.

Tách dữ liệu huấn luyện và kiểm tra: Sử dụng phương pháp train_test_split từ scikit-learn để tách dữ liệu thành tập huấn luyện và tập kiểm tra. Điều này giúp đánh giá hiệu suất của mô hình trên dữ liệu chưa từng được nhìn thấy trước đó.

Huấn luyện mô hình: Sử dụng thuật toán SVD dựa trên dữ liệu được chuẩn hóa như Bảng 2.

File kết quả của quá trình huấn luyện mô hình sẽ được sử dụng là CSDL cho máy chủ Python. Từ đó, xây dựng hệ thống APIs cho website bán hàng.

4.2. Quy trình xây dựng website bán hàng tích hợp hệ thống gợi ý

Xác định yêu cầu và mục tiêu: Đầu tiên, xác định rõ yêu cầu và mục tiêu của website bán hàng. Điều này bao gồm việc xác định loại sản phẩm, đối tượng khách hàng, chức năng và tính năng cần có trên website.

Xây dựng cơ sở dữ liệu sản phẩm: Xác định cấu trúc cơ sở dữ liệu để lưu trữ thông tin về sản phẩm, bao gồm mô tả, hình ảnh, giá cả, và các thuộc tính khác. Tạo các bảng dữ liệu và thiết lập quan hệ giữa chúng để có thể truy xuất dữ liệu hiệu quả.

Phát triển hệ thống gợi ý: Xây dựng hệ thống gợi ý dựa trên lịch sử người dùng. Sử dụng phân tích dữ liệu lịch sử của người dùng, áp dụng các phương pháp như lọc cộng tác, học máy hoặc học sâu để xác định các mục tiêu hoặc nội dung tương tự mà người dùng có thể quan tâm. Đảm bảo tích hợp hệ thống gợi ý vào website để hiển thị các sản phẩm gợi ý cho người dùng.

Triển khai và duy trì: Sau khi hoàn thiện và kiểm tra, triển khai website và hệ thống gợi ý lên môi trường sản xuất. Đồng thời, đảm bảo duy trì và cập nhật thường xuyên để tiếp tục nâng cao chất lượng và hiệu suất của website bán hàng và hệ thống gợi ý.

4. Kết luận

Trong quá trình làm website bán hàng thời trang Fashion shop và tích hợp hệ thống gợi ý, dự án đã đạt được những kết quả quan trọng sau:

Tăng trải nghiệm mua sắm: Bằng cách tích hợp hệ thống gợi ý, dự án đã cung cấp cho khách hàng những gợi ý sản phẩm phù hợp với sở thích và quan tâm cá nhân. Điều này giúp nâng cao trải nghiệm mua sắm của khách hàng, giúp họ tìm thấy các sản phẩm phù hợp nhanh chóng và thuận tiện.

Tăng doanh số bán hàng: Hệ thống gợi ý giúp khách hàng khám phá những sản phẩm mới và tương tự, khuyến nghị các sản phẩm phù hợp với lựa chọn của họ. Điều này có thể dẫn đến việc tăng cường khả năng chuyển đổi và tăng doanh số bán hàng cho Fashion shop.

Cải thiện sự đa dạng sản phẩm: Bằng cách gợi ý các sản phẩm tương tự hoặc phù hợp với lựa chọn của khách hàng, dự án đã giúp khách hàng khám phá được nhiều sản phẩm mới và đa dạng hơn. Điều này giúp nâng cao lựa chọn và tạo ra một trải nghiệm mua sắm đa dạng và thú vị.

Tối ưu hóa tiếp thị và quảng cáo: Hệ thống gợi ý cũng cung cấp thông tin về sở thích và hành vi của khách hàng, từ đó dự án có thể tối ưu hóa chiến dịch tiếp thị và quảng cáo. Điều này giúp chúng tôi hiểu rõ hơn về đối tượng khách hàng và tùy chỉnh các chiến lược tiếp thị để tăng hiệu quả và tối đa hóa tỷ lệ chuyển đổi.

Ngoài ra, các tính năng nên được xem xét và phát triển trong tương lai cho website bán hàng thời trang như là tích hợp tính năng đánh giá và nhận xét, mở rộng tính năng gợi ý, tối ưu hóa trải nghiệm di động, mở rộng phạm vi sản phẩm và thương hiệu, tích hợp tính năng thanh toán và vận chuyển, phát triển chiến lược tiếp thị và xây dựng cộng đồng,...

Tài liệu tham khảo

- [1] Vũ Hữu Tiệp (26/12/2016) “Giới thiệu về *Machine Learning*”.
- [2] Vũ Hữu Tiệp (24/05/2017) “*Neighborhood-based collaborative filtering*”.
- [3] Vũ Hữu Tiệp (07/06/2017) “*Singular value decomposition*”.
- [4] Đặng Hoàng Vũ (2016) “*Recommender systems*”
- [5] Blog Trí Tuệ Nhân Tạo (05/09/2020) “*Hệ thống gợi ý sản phẩm*”.
- [6] Yuanzhe Peng (31/05/2022) “*A survey on modern recommendation system based on big data*”.
- [7] Abhinav Ajitsaria (2019) “*Build a recommendation engine with collaborative filtering*” .
- [8] IBM Technology (14/07/2021) “*What is Machine Learning?*”
- [9] ProtonX (29/02/2020) “*Giới thiệu về Machine Learning + Deep Learning (Machine Learning + Deep Learning Introduction)*”.
- [10] Geeksforgeeks (31/03/2023) “*Machine Learning Tutorial*”.
- [11] Kaggle Competition (2022) “*H&M Personalized Fashion Recommendations*”
- [12] KHANHPHAMDINH (2018) “*Singular Value Decomposition*”