

XÂY DỰNG HỆ THỐNG GỢI Ý DỰA TRÊN HỒI QUY LASSO

Đàm Quang Viễn, Lư Thục Oanh
Trường Cao đẳng nghề Cần Thơ

Tóm tắt: Bài báo này đề xuất một mô hình hệ thống gợi ý mới sử dụng phương pháp hồi quy Lasso để tính toán độ tương đồng giữa người dùng hoặc giữa các sản phẩm, từ đó cải thiện độ chính xác của gợi ý. Phương pháp đề xuất xây dựng hai ma trận tương đồng riêng biệt – một ma trận từ dữ liệu đánh giá của người dùng và một ma trận từ các giá trị dự đoán của mô hình hồi quy Lasso – rồi kết hợp chúng để đưa ra đề xuất. Kết quả thí nghiệm trên các tập dữ liệu chuẩn cho thấy hệ thống gợi ý dựa trên Lasso đạt được độ chính xác cao hơn so với các phương pháp lọc cộng tác truyền thống. Các phép đo như Precision, Recall và đường cong ROC được sử dụng để đánh giá hiệu năng, khẳng định khả năng xử lý thiếu thông tin đánh giá của Lasso. Nghiên cứu nhấn mạnh vai trò quan trọng của hệ thống gợi ý trong bối cảnh thương mại điện tử phát triển và gợi mở hướng ứng dụng hồi quy Lasso trong lĩnh vực này.

Từ khóa: Hệ thống gợi ý, Hồi quy Lasso, Lọc cộng tác, Độ tương đồng, Xây dựng mô hình.

BUILDING A RECOMMENDATION SYSTEM BASED ON LASSO REGRESSION

Abstract: This paper presents a novel recommender system model leveraging Lasso regression to compute user/item similarities and enhance recommendation accuracy. The proposed approach constructs two similarity matrices – one from users' rating data and one from Lasso-based predictions – and integrates them for generating suggestions. Experimental evaluation on benchmark datasets demonstrates that the Lasso-based recommender achieves higher accuracy than conventional collaborative filtering models. Performance is measured using metrics such as precision, recall and ROC curves, confirming that Lasso regression effectively addresses sparse rating data issues. The study underscores the significance of recommender systems in e-commerce contexts and highlights the potential of Lasso regression for improving recommendation quality.

Keywords: Recommender System, Lasso Regression, Collaborative Filtering, Similarity Measures, Prediction Model.

Nhận bài: 02/03/2026

Phản biện: 21/03/2026

Duyệt đăng: 24/03/2026

I. ĐẶT VẤN ĐỀ

Trong thời đại số hiện nay, người dùng phải tiếp cận với một lượng lớn thông tin và tùy chọn. Hệ thống gợi ý (recommender system) ra đời như một công cụ cá nhân hóa quan trọng, giúp người dùng tìm kiếm sản phẩm hoặc nội dung phù hợp với sở thích của họ. Các hệ gợi ý được ứng dụng rộng rãi trong thương mại điện tử, mạng xã hội và giải trí. Ví dụ, Amazon và các nền tảng trực tuyến khác đã thành công trong việc sử dụng hệ thống gợi ý để đề xuất sản phẩm cho khách hàng. Việc xây dựng các mô hình gợi ý hiệu quả giúp tiết kiệm thời gian tìm kiếm thông tin cho người dùng và tăng doanh thu cho doanh nghiệp.

Trong lĩnh vực gợi ý cá nhân hóa, **lọc cộng tác (Collaborative Filtering – CF)** là một trong những phương pháp phổ biến nhất. Các mô hình CF dựa trên đánh giá của người dùng để đề xuất sản phẩm hoặc dịch vụ. Nhiều nghiên cứu đã chỉ ra CF là giải pháp hiệu quả để giải quyết “bài toán bùng nổ thông tin” trên hệ thống trực tuyến. Tuy nhiên, lọc cộng tác truyền thống vẫn gặp phải các hạn chế thực tế: dữ liệu đánh giá thường rất khan hiếm, đặc biệt khi xuất hiện người dùng mới hoặc sản phẩm mới (vấn đề cold start). Cụ thể, Huynh và cộng sự (2022) nhấn mạnh rằng thiếu thông tin đánh giá (do người dùng/mục mới hoặc ma trận đánh giá thưa) là thách thức lớn của các mô hình CF.

Để khắc phục những hạn chế này, nghiên cứu sử dụng **hồi quy Lasso (Least Absolute Shrinkage and Selection Operator)** như một giải pháp thay thế hoặc bổ sung cho CF. Lasso là một kỹ thuật hồi quy tuyến tính có phạt L1, giúp lựa chọn biến và thu nhỏ các hệ số về 0, từ đó cải thiện hiệu quả trên dữ liệu có chiều cao và bị thưa. Trong phương pháp đề xuất của Huynh et al. (2022), mô hình sử dụng hồi quy Lasso để tính toán độ tương đồng giữa người dùng/sản phẩm và kết hợp nó với ma trận tương đồng truyền thống. Kết quả thực nghiệm cho thấy hệ thống gợi ý dựa trên Lasso đạt độ chính xác cao hơn so với phương pháp lọc cộng tác thông thường. Phương pháp này tận dụng khả năng xử lý dữ liệu thiếu thông tin của Lasso để cải thiện khả năng khuyến nghị.

II. NỘI DUNG NGHIÊN CỨU

2.1. Mục tiêu nghiên cứu

1. Xây dựng một hệ thống gợi ý cá nhân hóa sử dụng mô hình hồi quy Lasso để tính toán độ tương đồng giữa người dùng hoặc giữa các sản phẩm, từ đó đề xuất mặt hàng phù hợp cho người dùng.

2. Thiết lập ma trận tương đồng dựa trên kết quả dự đoán của mô hình Lasso kết hợp với ma trận tương đồng từ dữ liệu đánh giá truyền thống, nhằm khai thác triệt để thông tin và khắc phục tình trạng dữ liệu thưa.

3. Đánh giá hiệu năng của hệ thống gợi ý mới trên các tập dữ liệu chuẩn (ví dụ: MovieLens) bằng các chỉ số đánh giá như RMSE, Precision, Recall và AUC-ROC, so sánh với phương pháp lọc cộng tác cổ điển để chứng minh hiệu quả của phương pháp đề xuất.

2.2. Đối tượng nghiên cứu

- **Dữ liệu nghiên cứu:** Các tập dữ liệu xếp hạng (rating matrix) từ người dùng đối với sản phẩm, ví dụ như tập MovieLens phổ biến trong nghiên cứu hệ gợi ý, được sử dụng để huấn luyện và kiểm thử hệ thống.

- **Mô hình gợi ý:** Hệ thống gợi ý dựa trên phương pháp lọc cộng tác kết hợp hồi quy Lasso. Đối tượng cụ thể là mô hình hồi quy tuyến tính có phạt L1 (Lasso) được áp dụng vào dữ liệu đánh giá.

- **Biến và chỉ tiêu:** Các biến đầu vào gồm điểm đánh giá của người dùng trên sản phẩm, các biến đầu ra là ma trận tương đồng giữa người dùng hoặc sản phẩm và kết quả đề xuất. Các chỉ tiêu chính để đánh giá bao gồm độ chính xác của dự đoán xếp hạng (ví dụ: RMSE) và độ hiệu quả của đề xuất (Precision, Recall, AUC-ROC).

2.3. Ý nghĩa khoa học và thực tiễn của nghiên cứu

- **Ý nghĩa khoa học:** Nghiên cứu này mở rộng ứng dụng của hồi quy Lasso trong lĩnh vực hệ thống gợi ý, một hướng nghiên cứu ít được quan tâm trước đây. Việc kết hợp Lasso để tính độ tương đồng cung cấp một góc tiếp cận mới cho bài toán lọc cộng tác khi đối mặt với dữ liệu đánh giá thưa. Kết quả phân tích chỉ ra rằng hệ thống gợi ý sử dụng Lasso cho độ chính xác cao hơn đáng kể so với các mô hình CF truyền thống, khẳng định hiệu quả của phạt L1 trong việc xử lý thiếu dữ liệu. Những đóng góp này đóng vai trò tham khảo cho cộng đồng nghiên cứu trong việc phát triển các thuật toán gợi ý cải tiến dựa trên mô hình hồi quy và học máy.

- **Ý nghĩa thực tiễn:** Hệ thống gợi ý chính xác cao giúp cải thiện trải nghiệm người dùng và tăng khả năng doanh thu cho doanh nghiệp thương mại điện tử. Như đã nêu, hệ thống gợi ý đang trở thành công cụ không thể thiếu trong các nền tảng trực tuyến như thương mại điện tử, mạng xã hội và giải trí. Áp dụng mô hình Lasso giúp giải quyết vấn đề dữ liệu đánh giá thiếu hụt hoặc không đầy đủ (cold-start), từ đó nâng cao hiệu quả khuyến nghị. Đặc biệt, trong bối cảnh Internet phát triển mạnh và lượng người dùng trực tuyến ngày càng

ngày càng nhiều, việc sử dụng Lasso có thể cải thiện độ ổn định và độ chính xác của các đề xuất gợi ý trong điều kiện thực tế.

2.4. Cơ sở lý thuyết

- **Hệ thống gợi ý:** Hệ thống gợi ý cá nhân hóa nhằm đề xuất các mục phù hợp dựa trên sở thích của người dùng. Cơ sở lý thuyết của hệ thống gợi ý bao gồm ba loại chính: dựa trên nội dung (content-based), lọc cộng tác (user-based/item-based collaborative filtering) và lai (hybrid). Lọc cộng tác là phương pháp phổ biến nhất, sử dụng ma trận đánh giá của người dùng để tính độ tương đồng. Nhiệm vụ chính của nó là khai thác mẫu quan hệ giữa người dùng và sản phẩm để dự đoán xếp hạng. Các phép đo thông dụng bao gồm hệ số tương quan Pearson, khoảng cách Cosine, v.v. Đánh giá hiệu năng hệ thống thường dựa trên các chỉ số như RMSE (Root Mean Square Error) cho độ chính xác dự đoán và Precision/Recall, AUC-ROC cho hiệu quả đề xuất. Hiện nay, các xu hướng mới như học sâu (deep learning) và đồ thị tri thức cũng được nghiên cứu để nâng cao khả năng của hệ thống gợi ý. Ví dụ, Zhang và cộng sự (2019) chỉ ra rằng phương pháp học sâu đã đạt được kết quả nổi bật trong lĩnh vực gợi ý nhờ khả năng học đặc trưng từ dữ liệu lớn. Tuy nhiên, những mô hình này thường yêu cầu dữ liệu khổng lồ và tính toán phức tạp.

- **Hồi quy Lasso:** Hồi quy Lasso là một biến thể của hồi quy tuyến tính với phần phạt L1, giúp lựa chọn biến và giảm hiện tượng quá khớp trong mô hình. Công thức tổng quát của Lasso gồm hàm mất mát bình phương cộng với điều khoản phạt L1 trên các hệ số. Đặc điểm của Lasso là có thể biến một số hệ số thành 0, do đó tự động chọn các biến quan trọng. Điều này đặc biệt hữu ích khi số biến đầu vào (số sản phẩm hoặc người dùng) lớn và dữ liệu bị khan hiếm. Khi áp dụng vào hệ gợi ý, Lasso cho phép dự đoán các giá trị đánh giá tiềm năng và sử dụng chúng để tính toán độ tương đồng giữa người dùng hoặc sản phẩm một cách hiệu quả. Theo Huynh et al. (2022), mô hình gợi ý dựa trên hai phương pháp tính tương đồng – một từ hồi quy Lasso và một từ dữ liệu đánh giá – đã được đề xuất để cải thiện độ chính xác của các đề xuất. Các lý thuyết thống kê liên quan như việc chọn tham số tối ưu (thông thường sử dụng cross-validation) và đánh giá tính ổn định của mô hình cũng được áp dụng khi huấn luyện Lasso.

- **Các khái niệm bổ sung:** Trong nghiên cứu

này, các thước đo Precision, Recall và đường cong ROC được sử dụng để đánh giá hiệu năng gợi ý. Precision và Recall đo lường độ chính xác và độ phủ của đề xuất, còn AUC-ROC đánh giá khả năng phân loại chính xác giữa các đề xuất đúng

sai. Những khái niệm này giúp đánh giá toàn diện hơn so với chỉ dùng sai số RMSE.

2.5. Phương pháp nghiên cứu

Mô hình hệ thống gợi ý dựa trên hồi quy Lasso được xây dựng dựa trên mô hình sau:



Trong đó:

- Dữ liệu xếp hạng $U \times I$ là dữ liệu xếp hạng người dùng cho các sản phẩm.

- Ma trận tương đồng dựa trên độ đo tương tự là ma trận tương đồng được tính dựa trên các độ đo tương tự như Cosin, Pearson, Jaccard...

0 Ma trận tương đồng dựa trên Lasso là ma trận dựa trên mô hình hồi quy Lasso.

Sau khi tính được các ma trận tương đồng việc xây dựng kết quả tư vấn sẽ được xây dựng dựa vào phương pháp lọc cộng tác dựa trên người dùng hoặc lọc cộng tác dựa trên sản phẩm và đưa ra kết quả tư vấn.

Ví dụ ứng dụng Lasso

Ma trận rating với 6 users và 8 item

Ma trận đánh giá với 6 người dùng giả định và 8 sản phẩm được đánh giá, các đánh giá theo mức đánh giá từ 1 đến 5 đây là tập dữ liệu nhỏ. Ở đây chúng tôi sẽ xây dựng hệ thống gợi ý dùng phương pháp lọc cộng tác dựa trên mối tương đồng của người dùng dựa thuật toán k láng giềng (K Nearest Neighbor - KNN) với $k=3$ dùng độ đo Cosine và độ đo Pearson để tìm k láng giềng gần nhất, sau đó dùng hồi quy Lasso để tìm k láng giềng để kiểm tra kết quả.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
	:	:	:	:	:	:	:	:
u_a	?	?	4.0	3.0	?	1.0	?	5.0

Nghiên cứu sử dụng mô hình hồi quy Lasso nhằm tìm ra mối tương đồng giữa các người dùng $u_k (k=\overline{1, n})$ hoặc các sản phẩm $i_k (k=\overline{1, p})$ với

người dùng $i_k (k=\overline{1, p})$ hoặc sản phẩm $u_j (j \neq k)$ từ đó tìm được ma trận tương đồng, từ tập dữ liệu

mới này chúng tôi xây dựng hệ thống gợi ý cho người dùng user ui có sử dụng sản phẩm ij dựa trên phương pháp lọc cộng tác.

Giải thuật lọc cộng tác trên người dùng (UBCF):

Tìm độ đo tương đồng giữa Ua các Users bằng độ đo Cosine

$$sim_{cosine}(\vec{X}, \vec{Y}) = \frac{\sum_{i=1}^n (X_i Y_i)}{\sqrt{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i^2}}$$

Ma trận thể hiện độ đo tương đồng khi sử dụng độ đo Cosine:

	u1	u2	u3	u4	u5	u6	ua
u1	0.0000000	0.18478729	0.3167782	0.1686869	0.3606679	0.5466082	0.52484882
u2	0.1847873	0.0000000	0.6000000	0.4694765	0.2927700	0.5070926	0.02366905
u3	0.3167782	0.6000000	0.0000000	0.8867889	0.5855400	0.5916080	0.61539539
u4	0.1686869	0.46947648	0.8867889	0.0000000	0.7126966	0.4629100	0.58338335
u5	0.3606679	0.29277002	0.5855400	0.7126966	0.0000000	0.5773503	0.40422604
u6	0.5466082	0.50709255	0.5916080	0.4629100	0.5773503	0.0000000	0.42008403
ua	0.5248488	0.02366905	0.6153954	0.5833834	0.4042260	0.4200840	0.00000000

Hay ta có độ đo tương đồng của ua với các users khác dùng độ đo Cosine là:

	u ₁	u ₂	u ₃	u ₄	u ₅	u ₆
ua	0.5248488	0.02366509	0.6153954	0.5833834	0.4042260	0.4200840

Tìm 3 láng giềng gần với ua là : u1 , u4 và u3.

Tìm độ đo tương đồng giữa Ua các Users bằng độ đo Pearson:

$$sim_{pearson}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Ma trận thể hiện độ đo tương đồng khi sử dụng độ đo Peason:

	u1	u2	u3	u4	u5	u6	ua
u1	0.0000000	0.5000000	0.5000000	0.5000000	0.5000000	0.5206449	0.5313457
u2	0.5000000	0.0000000	0.6035544	0.5284967	0.5000000	0.5594920	0.5000000
u3	0.5000000	0.6035544	0.0000000	0.7824521	0.5619759	0.5249112	0.5680024
u4	0.5000000	0.5284967	0.7824521	0.0000000	0.6579291	0.5000000	0.5512293
u5	0.5000000	0.5000000	0.5619759	0.6579291	0.0000000	0.5741184	0.5042105
u6	0.5206449	0.5594920	0.5249112	0.5000000	0.5741184	0.0000000	0.5000000
ua	0.5313457	0.5000000	0.5680024	0.5512293	0.5042105	0.5000000	0.0000000

Hay ta có độ đo tương đồng của ua với các users khác dùng độ đo Peason là:

	u ₁	u ₂	u ₃	u ₄	u ₅	u ₆
ua	0.5313475	0.5000000	0.5680024	0.5512293	0.5042105	0.5000000

Tìm 3 láng giềng gần nhất với ua là : u3 , u1 và u4

Tìm độ đo tương đồng giữa Ua các Users bằng mô hình hồi quy:

Xây dựng mô hình hồi quy với ua là biến phụ thuộc u_k ($k=1,6$) là các biến độc lập ta có mô hình hồi quy: $u_a = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 u_3 + \beta_4 u_4 + \beta_5 u_5 + \beta_6 u_6$

Với phương pháp bình phương tối thiểu ta tính được hệ số hồi quy như sau:

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-15.8806	9.7728	-1.625	0.351
u1	4.9403	2.7637	1.788	0.325
u2	0.4478	0.8064	0.555	0.677
u3	-2.4925	2.0543	-1.213	0.439
u4	7.9403	4.5838	1.732	0.333
u5	-9.0299	5.1499	-1.753	0.330
u6	5.0299	3.2067	1.569	0.361

	u ₁	u ₂	u ₃	u ₄	u ₅	u ₆
ua	4.9403	0.4478	-2.4925	7.9403	-9.0299	5.0299

Vậy phương trình hồi quy là:

$$u_a = -15.880 + 4.940u_1 + 0.4479u_2 - 2.492u_3 + 7.940u_4 - 9.029u_5 + 5.029u_6$$

Có thể nói các biến u₁ và u₄ và u₆ là các biến có tác động nhiều nhất đến biến phụ thuộc ua. với việc chọn hệ số hồi quy này chưa phù hợp với kết quả trên. Để điều chỉnh hệ số β ta dùng hồi quy Lasso.

Tìm độ đo tương đồng giữa Ua các Users bằng mô hình Lasso:

Xây dựng mô hình hồi quy Lasso tương tự như

$$\hat{\beta}^{lasso} = (y_1 - \beta_0 - (\beta_1x_{11} + \beta_2x_{12} + \beta_3x_{13} + \beta_4x_{14} + \beta_5x_{15} + \beta_6x_{16})^2 + (y_2 - \beta_0 - (\beta_2x_{21} + \beta_2x_{22} + \beta_3x_{23} + \beta_4x_{24} + \beta_5x_{25} + \beta_6x_{26})^2 + (y_3 - \beta_0 - (\beta_3x_{31} + \beta_3x_{33} + \beta_3x_{33} + \beta_4x_{34} + \beta_5x_{35} + \beta_6x_{36})^2 + (y_4 - \beta_0 - (\beta_4x_{41} + \beta_4x_{42} + \beta_3x_{43} + \beta_4x_{44} + \beta_5x_{45} + \beta_6x_{46})^2 + (y_5 - \beta_0 - (\beta_5x_{51} + \beta_5x_{52} + \beta_3x_{53} + \beta_4x_{55} + \beta_5x_{55} + \beta_6x_{56})^2 + \frac{1}{2}(y_6 - \beta_0 - (\beta_6x_{61} + \beta_6x_{62} + \beta_3x_{63} + \beta_4x_{65} + \beta_6x_{66} + \beta_6x_{66})^2 + \lambda(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5| + |\beta_6|)$$

Ta có thể tính được bằng cách giải hệ phương trình đạo hàm riêng hoặc dùng ma trận nghịch đảo, tuy nhiên có thể tìm λ được bằng cách chọn λ trong khoảng (0.0001;1000) chia thành 100 khoảng bằng nhau tuy nhiên với giá trị này hơi khó quan sát để thuận tiện trong biểu diễn ta dùng hàm log₁₀x với x ∈ (-3,3) sẽ thuận

mô hình hồi quy như trên ta có: với ua là biến phụ thuộc uk (k=(1,6)) là các biến độc lập ta có mô hình hồi quy:

$$u_a = \beta_0 + \beta_1u_1 + \beta_2u_2 + \beta_3u_3 + \beta_4u_4 + \beta_5u_5 + \beta_6u_6$$

Tìm hệ số hồi quy dựa vào cực tiểu hóa biểu thức:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1 \right) \text{ với } \|Y - X\beta\|_2^2 = \sum_{i=0}^n (y_i - (x\beta)_i)^2; \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \text{ và } \lambda > 0$$

tiện trong biểu diễn giá trị quan sát. Với hàm cv.glmnet ta có thể thực hiện khá thuận tiện, và với giá trị lambda nhỏ nhất đó ta có được các βi là ứng với những biến có tác động nhiều nhất đến biến phụ thuộc y.

Với một số giá trị lambda cụ thể ta có tập các hệ số hồi quy:

(Intercept)	0.9762704	0.53409427	-0.9180306	-2.2404440	-2.9376759	-3.1503351
u1	0.1756087	0.31067998	0.7202244	1.0931846	1.2878268	1.3443361
u2	-0.6356662	-0.60524243	-0.5121252	-0.4273300	-0.3888393	-0.3859779
u3	0.7080039	0.63856368	0.3616423	0.1094653	.	.
u4	0.0148687	0.21882910	0.9018762	1.5239063	1.8366240	1.9103757
u5	-0.2614310	-0.47791217	-1.2345390	-1.9235892	-2.2842397	-2.3906673
u6	.	0.05109732	0.4916134	0.8927886	1.0997992	1.1567136

Với giá trị lambda phù hợp ta có hệ số hồi quy:

	u ₁	u ₂	u ₃	u ₄	u ₅	u ₆
ua	0.1756087	-0.6356662	0.7080039	0.0148687	-0.2614310	0

Thấy có 3 biến u₃, u₁ và u₄ có tác động nhiều nhất đến biến phụ thuộc ua

Tính giá trị xếp hạng của người dùng ua dựa trên k láng giềng

Sau khi có được k láng giềng đã xác định như trên thì việc rating cho các sản phẩm chưa được đánh giá của người dùng ua được tính toán với công thức (2.1):

$$\hat{r}_{aj} = \frac{1}{|\mathcal{N}(a)|} \sum_{i \in \mathcal{N}(a)} r_{ij}$$

$$\hat{r}_{a1} = \frac{1}{2}(3.0 + 4.0) = 3.5,$$

$$\hat{r}_{a2} = \frac{1}{1}(4.0) = 4.0,$$

$$\hat{r}_{a5} = \frac{1}{3}(1.0 + 2.0 + 1.0) = 1.5$$

$$\hat{r}_{a7} = \frac{1}{1}(2.0) = 2.0$$

Các giá trị đã được rating được điền vào bảng ma trận như sau:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
	:	:	:	:	:	:	:	:
u_a	?	?	4.0	3.0	?	1.0	?	5.0
\hat{r}_a	3.5	4.0			1.3		2.0	

Kết quả đánh giá dựa trên ngưỡng dùng tương đồng

Vậy ta đã có bảng xếp hạng của người dùng a cho tất cả các sản phẩm i ($i = (1,8)$), với cách làm tương tự ta ta lấp đầy được ma trận tương đồng là đánh giá cho tất cả các sản phẩm của tất cả người

dùng và ma trận này được lưu trữ.

Giải thuật lọc cộng tác trên sản phẩm (IBCF):

Với dữ liệu đã cho trên ta có thể tính ma trận thể hiện độ đo tương đồng giữa các item khi sử dụng độ đo Cosine:

	1	2	3	4	5	6	7	8
1	0.00000000	0.03984095	0.00000000	0.5635445	0.7470179	0.6343350	0.6761234	0.60944940
2	0.03984095	0.00000000	0.6666667	0.3698001	0.2083333	0.6123724	0.0000000	0.06537205
3	0.00000000	0.6666667	0.0000000	0.6933752	0.1250000	0.6123724	0.0000000	0.49029034
4	0.56354451	0.36980013	0.6933752	0.0000000	0.3466876	0.8492078	0.3922323	0.87028527
5	0.74701788	0.20833333	0.1250000	0.3466876	0.0000000	0.6634035	0.1767767	0.26965969
6	0.63433505	0.61237244	0.6123724	0.8492078	0.6634035	0.0000000	0.2886751	0.64051262
7	0.67612340	0.00000000	0.0000000	0.3922323	0.1767767	0.2886751	0.0000000	0.55470020
8	0.60944940	0.06537205	0.4902903	0.8702853	0.2696597	0.6405126	0.5547002	0.00000000

Hoặc dữ liệu đã cho trên ta có thể tính ma trận thể hiện độ đo tương đồng giữa các item khi sử dụng độ đo Pearson:

	1	2	3	4	5	6	7	8
1	0.00000000	0.50000000	0.50000000	0.5233556	0.6775956	0.5155281	0.7263484	0.5537086
2	0.50000000	0.00000000	0.6818182	0.50000000	0.50000000	0.6073686	0.50000000	0.50000000
3	0.50000000	0.6818182	0.00000000	0.6774254	0.50000000	0.6073686	0.50000000	0.5466819
4	0.5233556	0.50000000	0.6774254	0.00000000	0.50000000	0.7232389	0.5495514	0.7786210
5	0.6775956	0.50000000	0.50000000	0.50000000	0.00000000	0.5620773	0.50000000	0.50000000
6	0.5155281	0.6073686	0.6073686	0.7232389	0.5620773	0.00000000	0.50000000	0.50000000
7	0.7263484	0.50000000	0.50000000	0.5495514	0.50000000	0.50000000	0.00000000	0.6414198
8	0.5537086	0.50000000	0.5466819	0.7786210	0.50000000	0.50000000	0.6414198	0.00000000

Hoặc ta có ma trận tương đương với ma trận 8x8 như trên sau:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	\hat{r}_a
i_1	—	0.1	0	0.3	0.2	0.4	0	0.1	—
i_2	0.1	—	0.8	0.9	0	0.2	0.1	0	2.0
i_3	0	0.8	—	0	0.4	0.1	0.3	0.5	4.6
i_4	0.3	0.9	0	—	0.1	0.1	0	0.2	3.2
i_5	0.2	0	0.4	0	—	0.1	0.2	0.1	—
i_6	0.4	0.2	0.1	0	0.1	—	0	0.1	2.8
i_7	0	0.1	0.3	0	0.2	0	—	0	4.0
i_8	0.1	0	0.5	0.2	0.1	0.1	0	—	—

u_a	2	?	?	?	4	?	?	5
-------	---	---	---	---	---	---	---	---

Hình 12- Kết quả đánh giá dựa trên sản phẩm tương đồng

Tính sự tương đồng của dựa vào k ($k=3$) sản phẩm tương đồng với sản phẩm tương đồng mà ta có các giá trị rating dựa vào công thức (2.2):

$$\hat{r}_{ai} = \frac{1}{\sum_{j \in S(i) \cap \{l: r_{aj} \neq ?\}} s_{ij}} \sum_{j \in S(i) \cap \{l: r_{aj} \neq ?\}} s_{ij} r_{uj}$$

$$\hat{r}_{a2} = \frac{1}{s_{21} + s_{25} + s_{28}} (s_{21}.ra1 + s_{25}.ra5 + s_{28}.ra8)$$

$$\hat{r}_{a2} = \frac{1}{0.1 + 0 + 0} (0.1 \times 2 + 0 \times 4 + 0 \times 5) = 2.0$$

$$\hat{r}_{a3} = \frac{1}{0 + 0.4 + 0.5} (0 \times 2 + 0.4 \times 4 + 0.5 \times 5) = 4.6$$

$$\hat{r}_{a4} = \frac{1}{0.3 + 0 + 0.2} (0.3 \times 2 + 0 \times 4 + 0.2 \times 5) = 3.2$$

$$\hat{r}_{a6} = \frac{1}{0.4 + 0.1 + 0.1} (0.4 \times 2 + 0.1 \times 4 + 0.1 \times 5) = 2.8$$

$$\hat{r}_{a7} = \frac{1}{0 + 0.2 + 0} (0 \times 2 + 0.2 \times 4 + 0 \times 5) = 4.0$$

Tương tự ta cũng có bảng rating cho toàn bộ sản phẩm chưa được đánh giá

III. KẾT LUẬN

Hệ thống gợi ý có vai trò quan trọng trong thời gian gần đây đặc biệt là trong thời gian dịch Covid việc mua hàng, quảng cáo sản phẩm, thanh toán đều diễn ra với các hình thức online vì vậy có rất nhiều phương pháp nhằm tìm ra hệ thống gợi ý, cải tiến hệ thống gợi ý tạo ra các hướng nghiên cứu khác nhau cho các hệ tư vấn. Qua một thời gian nghiên cứu lý thuyết kết hợp với phần thực

nghiệm, luận văn tốt nghiệp “**Hệ thống gợi ý với hồi quy Lasso**” đã hoàn thành và đạt được những kết quả khả quan.

Về mặt lý thuyết: Hiểu rõ hệ thống gợi ý là gì và cách thức hoạt động của một hệ thống gợi ý. Khi áp dụng hệ thống gợi ý vào thực tiễn, các vấn đề cần quan tâm: nghiên cứu về độ đo và các mô hình hồi quy sử dụng trong hệ thống gợi ý, các phương pháp gợi ý dựa trên lọc cộng tác, so sánh và đánh giá hệ thống gợi ý như thế nào. Ngoài ra, luận văn áp dụng mô hình hồi quy Lasso, giải thuật k láng giềng và lý thuyết về các công thức, đánh giá mô hình hệ thống gợi ý.

Về mặt thực tiễn: Nghiên cứu xây dựng được hệ thống gợi ý trên ngôn ngữ R nhằm mục đích xem xét cải thiện hiệu quả của hệ thống gợi ý bằng cách kết hợp hệ thống gợi ý theo mô hình lọc cộng tác người dùng dựa trên mô hình hồi quy Lasso và lọc cộng tác trên người dùng và trên sản phẩm, từ đó so sánh, đánh giá các kết quả dựa trên các chỉ số lỗi (RMSE, MSE, MAE). Thực nghiệm trên tập dữ liệu phi nhị phân Movielens và đạt được kết quả tư vấn cuối cùng. Luận văn cũng đã sử dụng các phương pháp đánh giá hệ tư vấn để có các biểu đồ kết quả.

TÀI LIỆU THAM KHẢO

- G. Adomavicius, A. Tuzhilin, “Context-Aware Recommender System”, *Recommender System Handbook*, Springer, pp. 191–226, 2015.
- B. Amer, J. Alsakran, “An Automated Recommender System for Course Selection”, *Vol. 7- No. 3*, International Journal of Advanced Computer Science and Applications, pp. 1-10, 2016.
- Anind K. Dey, “Providing Architectural Support for Building Context Aware Applications”, *PhD thesis*, Georgia Institute of Technology, pp.1-34, 2000.
- Ariely D., Lynch, J., and Aparicio M., Learning by Collaborative and Individual-Based Recommendation Agents. *Journal of Consumer Psychology*, 14(1 & 2), 81–95, 2004.
- K. Bauman, A. Tuzhilin, “Location-Based Recommendation Systems”, *Encyclopedia of GIS*, Springer, pp. 43-92, 2017.
- Peter Buhlmann, Sara van de Geer, *Statistics for High-Dimensional Data*, Springer, 2011.