HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Trần Thị Nghĩa

NGHIÊN CỨU MỘT SỐ ĐỘ ĐO TƯƠNG TỰ CHO TƯ VẤN LỌC CỘNG TÁC

LUẬN VĂN THẠC SỸ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI – 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Trần Thị Nghĩa

NGHIÊN CỨU MỘT SỐ ĐỘ ĐO TƯƠNG TỰ CHO TƯ VẤN LỌC CỘNG TÁC

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

LUẬN VĂN THẠC SỸ KỸ THUẬT

(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. TRẦN ĐÌNH QUẾ

HÀ NỘI – 2022

LỜI CAM ĐOAN

Tôi cam đoan luận văn đề tài "***Nghiên cứu một số độ đo tương tự cho tư vấn lọc cộng tác***" là công trình nghiên cứu của riêng tôi.

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

Trần Thị Nghĩa

LỜI CẢM ƠN

Trong suốt quá trình thực hiện đề tài luận văn "Nghiên cứu một số độ đo tương tự cho tư vấn lọc cộng tác" tôi đã nhận được rất nhiều sự giúp đỡ, động viên tạo điều kiện từ thầy cô, gia đình và bạn bè. Tôi xin bày tỏ lòng cảm ơn chân thành về sự giúp đỡ động viên này.

Trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc tới PGS.TS. Trần Đình Quế - người đã định hướng cho tôi trong việc lựa chọn đề tài, đưa ra những nhận xét quý giá và trực tiếp hướng dẫn tôi trong suốt quá trình nghiên cứu và hoàn thiện luận văn.

Tiếp theo, tôi xin gửi lời cảm ơn chân thành tới tất cả các quý thầy cô giáo của Học viện Công nghệ Bưu chính Viễn thông đã giảng dạy và hướng dẫn cho tôi trong trong suốt quá trình học tập tại trường.

Cuối cùng, tôi xin bày tỏ lòng biết ơn chân thành đối với gia đình và bạn bè - những người luôn ở bên cạnh động viên, ủng hộ, cổ vũ và tạo điều kiện cho tôi hoàn thành khóa luận này.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
CF	Collaborative filtering	Lọc cộng tác
SVD	Singular Value Decomposition	Phương pháp phân tích suy biến
LIS	Latent Semantic Indexing	Lập chỉ mục ngữ nghĩa tiềm ẩn
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Phân cụm không gian dựa trên mật độ các ứng dụng với nhiễu
OPTICS	Ordering points to identify the clustering structure	Thuật toán phân cụm dựa vào thứ tự các điểm
BRICH	Balanced iterative reducing and clustering using hierarchies	Thuật toán giảm lặp và phân cụm cân bằng bằng cách sử dụng phân cấp
MAE	Mean-Absolute Error	Sai số tuyệt đối trung bình
RMSE	Root Mean Square Error	Sai số trung bình bình phương
MAP	Mean Average Precision	Độ chính xác trung bình tuyệt đối
COS	Cosine similarity	Tương tự Cosine
J	Jaccard index	Chỉ số Jaccard
E	Euclide distance	Khoảng cách Euclide
PCC	Pearson Correlation Coefficient	Hệ số tương quan Pearson
CPCC	Constrained Pearson Correlation	Hệ số tương quan Pearson ràng buộc
SPCC	Sigmoid Function-Based Pearson Correlation	Tương quan Pearson dựa trên chức năng Sigmoid

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

I. MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại phát triển của công nghệ thông tin như hiện nay việc lựa chọn các thông tin hữu ích là một vấn đề khó khăn với người dùng do có một sự gia tăng rất lớn về lượng thông tin có sẵn trên Web. Sự gia tăng to lớn này trong thông tin không thể xử lý dễ dàng được dẫn đến việc quá tải thông tin. Trong cuộc sống hàng ngày, mọi người thường dựa vào những khuyến nghị của người khác để lựa chọn thông tin thông qua lời nói, thư tham khảo, các tin tức từ các phương tiện truyền thông, hay từ những khảo sát chung…, hệ thống tư vấn (*Recommender systems*) hỗ trợ và tăng cường quá trình xã hội tự nhiên này để giúp người dùng sàng lọc thông tin bằng cách dự đoán và cung cấp cho người dùng một danh sách những cuốn sách, bài báo, trang web, phim ảnh, âm nhạc, nhà hàng, sản phẩm,…có thông tin thú vị và có giá trị nhất mà người dùng có khả năng quan tâm đến. Hiện nay nhiều trang thương mại đã được sử dụng hệ tư vấn rất thành công như hệ thống của Netflix, Amazon, Youtube...[16]

Lọc cộng tác (CF) là một phương pháp tiếp cận được sử dụng để đưa ra các đề xuất dựa trên mối tương quan giữa các tùy chọn của người dùng. Những lựa chọn này được tìm thấy bằng cách sử dụng các độ đo tương tự như: Hệ số tương quan Pearson, Tương quan Pearson hạn chế, Cosine, Jaccard, v.v. Vì lý do đó trong luận văn này tác giả sẽ nghiên cứu một số độ đo tương tự sử dụng cho tư vấn lọc cộng tác, sử dụng thuật toán K-means để phân tích và đánh giá hiệu quả của các độ đo tương tự. Có rất nhiều độ đo tương tự sử dụng trong các kỹ thuật lọc cộng tác như [3]: Tương tự Cosine (Cosine similarity), tương tự Cosine điều chỉnh (Adjusted Cosine Vector), hệ số tương quan Pearson (Pearson Correlation Coefficient), thông tin tương hỗ điều chỉnh (Adjusted Mutual Information), chỉ số Rand điều chỉnh (Adjusted Rank index), hệ số tương quan thứ tự bậc Spearman (Spearman rank-order correlation coefficient), tương tự Heuristic (Heuristic similarity), chỉ số Jaccard (Jaccard index), khoảng cách Euclide (Euclide distance), khoảng cách Manhattan (Manhattan distance),

khoảng cách Chebyshev (Chebyshev distance), độ tương tự tam giác (Triangle similarity), PCC có trọng số với RPB (improved PCC weighted with RPB),… Tuy nhiên trong luận văn này tác giả sẽ tập trung nghiên cứu một số độ đo tương tự như: Tương tự Cosine, hệ số tương quan Pearson, hệ số tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, chỉ số Jaccard, khoảng cách Euclide.

2. Tổng quan về vấn đề nghiên cứu

Hệ thống tư vấn được xây dựng dựa theo một trong hai mô hình chính đó là phương pháp lọc dựa trên nội dung và phương pháp lọc cộng tác. Kỹ thuật lọc dựa trên nội dung được thực hiện dựa vào việc so sánh các nội dung của thông tin hay những mô tả của hàng hoá để tìm ra những sản phẩm có sự tương đồng với những nhu cầu mà người dùng quan tâm trước đó. Kỹ thuật lọc theo nội dung được phát triển dựa vào việc kế thừa các phương pháp trích chọn đặc trưng trong lĩnh vực truy vấn thông tin. Để đưa ra một tập các đặc trưng phù hợp và đầy đủ, nội dung của tài liệu phải được biểu diễn dưới dạng hợp lý để máy tính có thể tự động tính toán, phân tích trọng số các đặc trưng của nội dung. Phương pháp này sẽ khó áp dụng trong những trường hợp trích chọn đặc trưng những nội dung phức tạp như dữ liệu đa phương tiện (hình ảnh, âm thanh, dịch vụ). Khác với lọc theo nội dung, lọc cộng tác chỉ sử dụng dữ liệu xếp hạng của người dùng để đưa ra dự đoán và đề xuất. Do đó, lọc cộng tác có thể lọc hiệu quả hơn trên nhiều sản phẩm khác nhau như phim, ảnh, âm thanh, hàng hoá. Mục đích của phương pháp tư vấn dựa trên lọc cộng tác là dự đoán một sản phẩm phù hợp cho người dùng hoặc dự đoán những sản phẩm mới dựa trên những sở thích trước đây hoặc những sở thích tương tự của những người dùng khác. Trong tư vấn lọc cộng tác được chia làm các kỹ thuật lọc khác nhau đó là: *Kỹ thuật lọc cộng tác dựa trên bộ nhớ* và *Kỹ thuật lọc cộng tác dựa trên mô hình*.

Kỹ thuật lọc cộng tác dựa trên bộ nhớ là một phương pháp tính toán mức độ giống nhau giữa người dùng này với người dùng khác hoặc sản phẩm này với sản phẩm khác sử dụng những dữ liệu trước đó của người dùng đã đánh giá.

Kỹ thuật lọc cộng tác dựa trên mô hình: Việc thiết kế và phát triển các mô hình (chẳng hạn như học máy, thuật toán khai thác dữ liệu) có thể cho phép hệ thống học cách nhận ra các mẫu phức tạp dựa trên dữ liệu đào tạo và sau đó đưa ra dự đoán thông minh cho các tác vụ lọc cộng tác đối với dữ liệu thử nghiệm hoặc dữ liệu trong thế giới thực dựa trên các mô hình đã học. Các thuật toán lọc cộng tác dựa trên mô hình, chẳng hạn như mô hình Bayes, mô hình phân cụm và mạng phụ thuộc, …

Để tính toán được mức độ giống nhau thì các độ đo tương tự đóng vai trò rất quan trọng. Trong kỹ thuật lọc cộng tác sử dụng các độ đo tương tự như [3]:

- *Hệ số tương quan Pearson*: Hệ số tương quan của Pearson là thống kê kiểm định đo lường mối quan hệ thống kê, hay sự liên kết giữa hai biến liên tục. Hệ số tương quan Pearson được biết đến là một phương pháp đo lường tốt nhất để đo lường mối liên hệ giữa các biến quan tâm vì nó dựa trên phương pháp hiệp phương sai. Nó cung cấp các thông tin về độ lớn của mối liên kết, mối tương quan, cũng như hướng của mối quan hệ.
- *Chỉ số Jaccard*: (hệ số tương tự Jaccard) là thước đo mức độ giống nhau của hai bộ dữ liệu, với phạm vi từ 0% đến 100%. Tỷ lệ phần trăm càng cao thì hai bộ dữ liệu càng giống nhau.
- *Tương tự Cosine*: là một phép đo được sử dụng để đo lường mức độ giống nhau giữa hai hoặc nhiều véc tơ.
- *Hệ số tương quan Pearson ràng buộc*: là một biến thể của hệ số tương quan Pearson phù hợp hơn khi có xếp hạng âm trong tập dữ liệu.
- *Tương quan Pearson dựa trên chức năng Sigmoid*: là sử dụng một hàm Sigmoid để giảm giá trị tương tự giữa các sản phẩm mà ít người dùng đã xếp hạng cả hai sản phẩm.
- *Khoảng cách Euclide*: Khoảng cách Euclide giữa hai điểm trong không gian Euclid được định nghĩa là độ dài của đoạn thẳng giữa hai điểm. Vì khoảng cách Euclide có thể được tìm thấy bằng cách sử dụng các điểm tọa độ và định lý Pythagoras, nó đôi khi được gọi là khoảng cách Pitago.

3. Mục đích nghiên cứu

Mục tiêu đặt ra của luận văn trong đề tài này là: Khảo sát các cách tiếp cận tư vấn lọc cộng tác bằng cách nghiên cứu một số độ đo tương tự sử dụng trong tư vấn lọc cộng tác, dùng thuật toán K-Means thử nghiệm và đánh giá các độ đo tương tự được sử dụng trong tư vấn lọc cộng tác.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Đề tài tập trung nghiên cứu các độ đo tương tự sử dụng cho tư vấn lọc cộng tác.

Phạm vi nghiên cứu: Sử dụng cho việc đánh giá hiệu quả của các độ đo tương tự sử dụng cho tư vấn lọc cộng tác.

5. Phương pháp nghiên cứu

Nghiên cứu lý thuyết về tư vấn lọc cộng tác và các độ đo tương tự bằng cách đọc và phân tích các tài liệu, công trình nghiên cứu đã được đăng tải.

Thử nghiệm và đánh giá các độ đo tương tự dựa trên bộ dữ liệu *MovieLens* trên trang web *https://grouplens.org/datasets/movielens/*

Chương 1. TỔNG QUAN VỀ TƯ VẤN LỌC CỘNG TÁC

1.1. Giới thiệu chung

Trong thời đại phát triển của công nghệ thông tin như hiện nay, các trang thương mại điện tử cung cấp lên đến hàng triệu các sản phẩm được bán. Lựa chọn giữa rất nhiều sản phẩm trở thành một công việc đầy thách thức đối với khách hàng. Hệ thống khuyến nghị xuất hiện để giải quyết vấn đề này. Hệ thống giới thiệu được sử dụng để cung cấp các khuyến nghị chất lượng giúp hướng dẫn khách hàng đưa ra quyết định mua sản phẩm. Hệ thống giới thiệu về cơ bản được sử dụng trong các trang web thương mại điện tử, trong đó đầu vào của hệ thống sẽ là phân tích hành vi mua của khách hàng được sử dụng để đưa ra danh sách giới thiệu các sản phẩm. Hệ thống giới thiệu đã thay đổi cách mọi người tìm kiếm sản phẩm, thông tin và mục tiêu chính của hệ thống giới thiệu là cung cấp cho khách hàng các khuyến nghị chính xác và chất lượng tốt. Hầu hết tất cả các hệ thống giới thiệu thường bắt đầu bằng cách tìm kiếm nhóm khách hàng đã mua hoặc xếp hạng các sản phẩm tương tự và trùng lặp với các sản phẩm đã mua hoặc xếp hạng của người dùng hiện tại. Có nhiều cách triển khai hệ thống khuyến nghị dựa trên nhiều yếu tố khác nhau và được áp dụng cho các bối cảnh khác nhau như hệ thống khuyến nghị bán lẻ lấy cảm hứng từ sinh học, tối ưu hóa siêu thông số cho hệ thống khuyến nghị hoặc hệ thống khuyến nghị dựa trên sự tương đồng về ngữ nghĩa. Một trong những công nghệ khuyến nghị đầu tiên và được sử dụng rộng rãi là "*Lọc cộng tác*" [15]. Lọc cộng tác (CF) là lọc thông tin tư vấn cho người dùng dựa trên tập hợp các hồ sơ người dùng có cùng sở thích. Các thuật toán lọc cộng tác được phân loại là dựa trên người dùng và dựa trên các sản phẩm. Hệ thống đề xuất cần lưu trữ thông tin về các tùy chọn của người dùng được gọi là hồ sơ người dùng. Hồ sơ của người dùng có thể được thu thập một cách rõ ràng hoặc ẩn ý. Người ta có thể yêu cầu người dùng đánh giá một cách rõ ràng những gì họ đã quan tâm đến. Một hồ sơ như vậy được điền rõ ràng bởi xếp hạng của người dùng. Hồ sơ ngầm dựa trên quan sát thụ động và chứa dữ liệu tương tác lịch sử của người dùng. Dựa trên chiến lược này, các sản phẩm mà người dùng khác đã quan tâm

tương tự với người dùng mục tiêu được đề xuất. Sự tương đồng giữa hai người dùng được tính toán với sự trợ giúp của xếp hạng do những người dùng khác thực hiện.

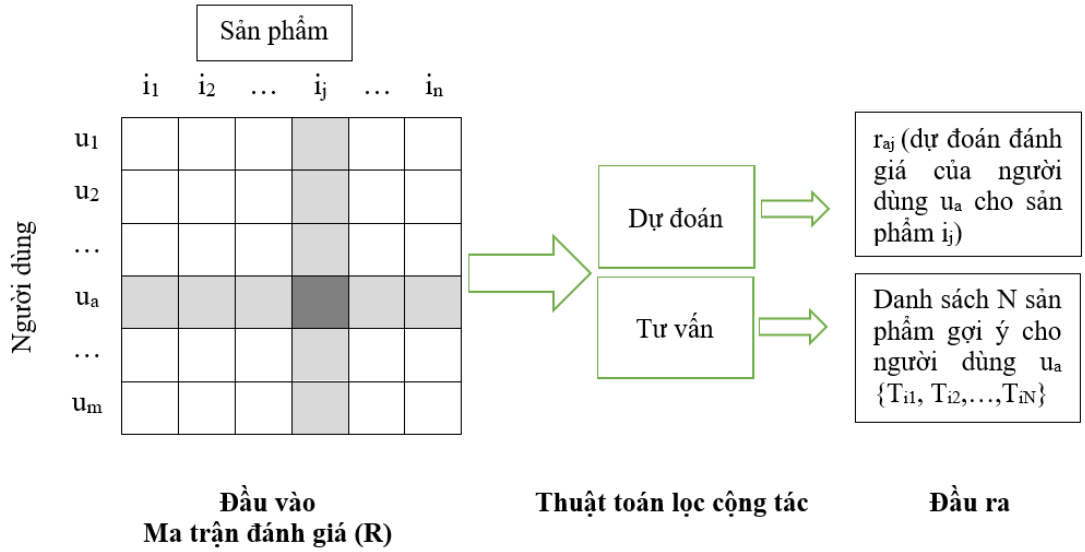
Thuật ngữ "*Collaborative filtering*" lần đầu tiên được Goldberg áp dụng cho hệ thống tư vấn Tapestry, kể từ đó *CF* đã trở thành một trong những kỹ thuật được sử dụng rộng rãi nhất để cung cấp các khuyến nghị dịch vụ cho người dùng trực tuyến [3]. Tuy nhiên, hệ thống tư vấn cho các cộng đồng lớn không thể phụ thuộc vào việc mỗi người biết những người khác. Sau đó, một số hệ thống giới thiệu tự động dựa trên xếp hạng đã được phát triển. Hệ thống nghiên cứu GroupLens cung cấp giải pháp lọc cộng tác bằng bút danh cho tin tức và phim trên Usenet. Ringo và Video Recommender là các hệ thống dựa trên web và email tạo ra các đề xuất về nhạc và phim tương ứng. Một số đặc biệt về truyền thông của ACM trình bày một số hệ thống khuyến nghị khác nhau [2].

Lọc cộng tác đã rất thành công trong cả thực tiễn tìm kiếm lại, trong cả ứng dụng thu thập thông tin và ứng dụng thương mại điện tử [2].

1.2. Bài toán lọc cộng tác

Mục tiêu của thuật toán lọc cộng tác là đề xuất các sản phẩm mới hoặc dự đoán tiện ích của một sản phẩm nhất định đối với người dùng cụ thể dựa trên sở thích trước đây của người dùng và ý kiến của những người dùng cùng chí hướng. Trong một kịch bản CF cổ điển có m là một danh sách người dùng ký hiệu là $U = \{u_1, u_2, \ldots, u_m\}$ và n là một danh sách các sản phẩm mà người dùng có thể lựa chọn ký hiệu là $I = \{i_1, i_2, .., i_n\}$. Mỗi người dùng u_i có một danh sách các sản phẩm mà người dùng đã đánh giá về sản phẩm đó gọi là S_u, mỗi sản phẩm $i_j \in I$ có thể là hàng hóa, phim, ảnh, tạp chí, tài liệu, sách, báo, dịch vụ hoặc bất kỳ dạng thông tin nào mà người dùng cần đến. Tiếp theo, ký hiệu $R=\{ r_{ij} \}$, $i = 1\ldots m$, $j = 1\ldots n$ là ma trận đánh giá, trong đó mỗi người dùng $u_i \in U$ đưa ra đánh giá của mình về một số sản phẩm $i_j \in I$ bằng một số r_{ij}. Mức độ ưa thích của người dùng u_i đối với sản phẩm i_j được thể hiện qua giá trị r_{ij}. Giá trị r_{ij} có thể được thu thập trực tiếp bằng cách dựa trên điểm xếp hạng hoặc thu thập gián tiếp xuất phát từ hồ sơ mua hàng, bằng cách phân tích nhật ký, bằng cách

khai thác các siêu liên kết web,....Nếu trong trường hợp người dùng u_i chưa biết đến sản phẩm i_j hoặc chưa đánh giá sản phẩm đó thì giá trị $r_{ij} = \emptyset$. Với một người dùng $u_a \in U$ (được gọi là người dùng đang hoạt động, người dùng cần được tư vấn, hay người dùng mục tiêu) nhiệm vụ của bài toán lọc cộng tác được thể hiện trong hình 1.1.



Hình 1.1: Sơ đồ thể hiện quy trình của hệ thống tư vấn lọc cộng tác

Ma trận đánh giá $R = (r_{ij})$ là thông tin đầu vào duy nhất của các kỹ thuật lọc cộng tác. Dựa trên ma trận đánh giá, các kỹ thuật lọc cộng tác thực hiện hai nhiệm vụ chính đó là dự đoán và tư vấn.

Dự đoán là một giá trị số r_{aj} thể hiện sở thích của người dùng u_a đối với những sản phẩm mà u_a chưa đánh giá ($r_{aj} = \emptyset$), trên cơ sở đó tư vấn cho u_a những sản phẩm được đánh giá cao.

Đề xuất một danh sách N sản phẩm mà người dùng u_a thích nhất.

Ví dụ về ma trận đánh giá $R = (r_{ij})$ trong hệ tư vấn lọc cộng tác gồm bốn người dùng trong tập $U = \{u_1, u_2, u_3, u_4\}$ đánh giá bảy sản phẩm trong tập $I = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$. Mỗi người dùng đều đưa ra các đánh giá của mình về các sản phẩm theo thang bậc $\{\emptyset, 1, 2, 3, 4, 5\}$.

Bảng 1.1: Ví dụ về ma trận đánh giá của lọc cộng tác

Người dùng	Sản phẩm						
	i_1	i_2	i_3	i_4	i_5	i_6	i_7
u_1	4			5	1		
u_2	5	5	4		4		5
u_3				2	4	5	
u_4	2	3					3

1.3. Đặc điểm và thách thức của lọc cộng tác

Việc vận dụng các thuật toán lọc cộng tác trong thương mại điện tử thường gặp phải rất nhiều vấn đề thách thức, đặc biệt là đối với các hệ thống mua sắm trực tuyến lớn như *eBay* và *Amazon.* Thông thường, một hệ thống giới thiệu cung cấp các khuyến nghị nhanh chóng và chính xác sẽ thu hút sự quan tâm của khách hàng và mang lại lợi ích cho các công ty. Đối với các hệ thống *CF,* việc đưa ra các dự đoán hoặc khuyến nghị đủ tiêu chuẩn phụ thuộc vào mức độ chúng giải quyết các thách thức, đó cũng là đặc điểm của các nhiệm vụ CF.

1.3.1. Dữ liệu thưa thớt

Trong thực tế, nhiều hệ thống khuyến nghị thương mại được sử dụng để đánh giá các bộ sản phẩm rất lớn. Do đó, ma trận đánh giá của người dùng được sử dụng để lọc cộng tác sẽ cực kỳ thưa thớt và hiệu suất của các dự đoán hoặc khuyến nghị của các hệ thống lọc cộng tác bị thách thức.

Thách thức về dữ liệu thưa xuất hiện trong một số tình huống như vấn đề *xuất phát nguội* xảy ra khi một người dùng mới được thêm vào hệ thống sẽ chưa có đánh giá sản phẩm nào hoặc chưa có các hành vi nào với sản phẩm mới được thêm vào sẽ chưa được người dùng nào đánh giá hoặc chưa được ai mua, xem, tìm kiếm. Rất khó để tìm thấy những người dùng hoặc sản phẩm tương tự vì không có đủ thông tin.

Để giảm bớt vấn đề dữ liệu thưa, nhiều cách tiếp cận đã được đề xuất ví dụ như các kỹ thuật giảm kích thước, chẳng hạn như kỹ thuật giảm số chiều (SVD), loại bỏ người dùng hoặc sản phẩm không đại diện hoặc không đáng kể để giảm kích thước của ma trận sản phẩm người dùng trực tiếp.

1.3.2. Khả năng mở rộng

Số lượng người dùng và sản phẩm tăng lên rất nhiều theo thời gian, do đó các thuật toán CF truyền thống sẽ phải giải quyết các vấn đề nghiêm trọng về khả năng mở rộng khi tài nguyên tính toán vượt quá mức thực tế hoặc mức có thể chấp nhận được. Ví dụ, với hàng chục triệu khách hàng (M) và hàng triệu danh mục riêng biệt (N), một thuật toán CF với độ phức tạp của O(n) đã quá lớn. Ngoài ra, nhiều hệ thống cần phản ứng ngay lập tức với các yêu cầu trực tuyến và đưa ra khuyến nghị cho tất cả người dùng bất kể lịch sử mua hàng và xếp hạng của họ, điều này đòi hỏi khả năng mở rộng cao của hệ thống CF.

Các kỹ thuật giảm kích thước như SVD có thể giải quyết vấn đề về khả năng mở rộng và nhanh chóng đưa ra các đề xuất chất lượng tốt, nhưng chúng phải trải qua các bước phân tích nhân tử ma trận tốn kém. Một thuật toán lọc cộng tác sử dụng SVD gia tăng tính toán trước quá trình phân rã bằng cách sử dụng những người dùng hiện có. Khi có các xếp hạng mới được thêm vào cơ sở dữ liệu, thuật toán sử dụng kỹ thuật chiếu gấp khúc để xây dựng một hệ thống gia tăng mà không cần tính toán lại mô hình chiếu từ đầu. Do đó, nó làm cho hệ thống giới thiệu có khả năng mở rộng cao.

1.3.3. Từ đồng nghĩa

Từ đồng nghĩa đề cập đến xu hướng của một số nội dung giống nhau nhưng có tên nhập khác nhau. Đa số các hệ thống tư vấn không thể phát hiện ra mối liên quan tiềm ẩn này do đó sẽ xử lý các sản phẩm này một cách khác biệt. Ví dụ: các nội dung có vẻ khác nhau như "*children movie*" và "*children film*" nhưng trên thực tế là cùng một nội dung, tuy nhiên các hệ thống CF dựa trên bộ nhớ sẽ không tìm thấy sự phù hợp nào giữa chúng để tính toán sự giống nhau. Thật vậy, mức độ thay đổi trong cách sử dụng thuật ngữ mô tả lớn hơn mức thường được nghi ngờ. Các từ đồng nghĩa sẽ làm giảm hiệu suất khuyến nghị của các hệ thống CF.

Những nỗ lực trước đây để giải quyết vấn đề đồng nghĩa phụ thuộc vào việc mở rộng thuật ngữ trí tuệ hoặc tự động, hoặc việc xây dựng một từ điển đồng nghĩa. Hạn chế đối với các phương pháp hoàn toàn tự động là một số thuật ngữ được bổ sung có thể có ý nghĩa khác với dự định, do đó dẫn đến sự suy giảm nhanh chóng của hiệu suất khuyến nghị.

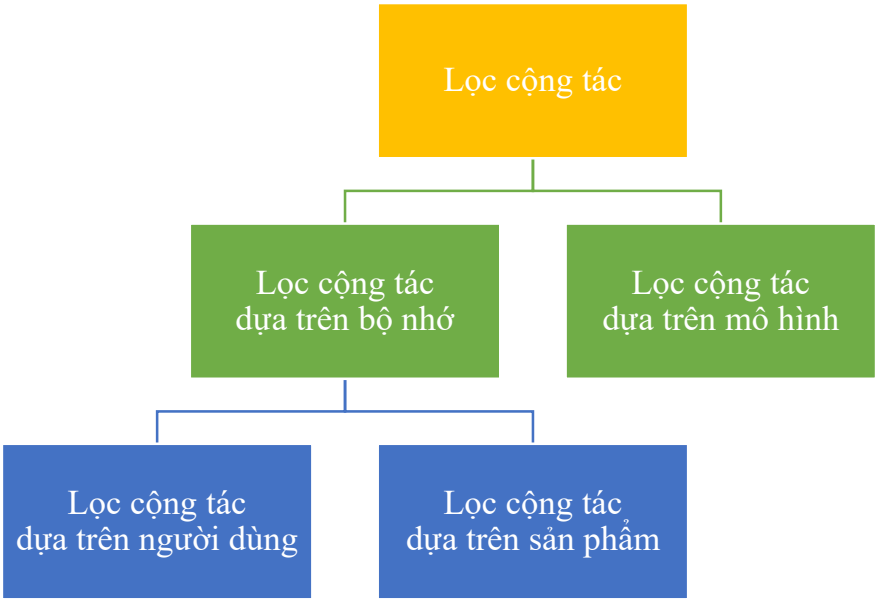
Các kỹ thuật SVD, đặc biệt là phương pháp lập chỉ mục ngữ nghĩa tiềm ẩn (LSI) có khả năng giải quyết các vấn đề đồng nghĩa. SVD lấy một ma trận lớn dữ liệu liên kết tài liệu thuật ngữ và xây dựng một không gian ngữ nghĩa nơi các thuật ngữ và tài liệu liên kết chặt chẽ được đặt gần nhau. SVD cho phép sắp xếp không gian để phản ánh các mẫu liên kết chính trong dữ liệu và bỏ qua các mẫu nhỏ hơn, ít quan trọng hơn. Hiệu suất của LSI trong việc giải quyết vấn đề đồng nghĩa là rất ấn tượng ở các mức thu hồi cao hơn, nơi độ chính xác thường khá thấp, do đó đại diện cho các cải tiến tỷ lệ lớn. Tuy nhiên, hiệu suất của phương pháp LSI ở mức thu hồi thấp nhất là kém.

1.3.4. Gray sheep và Black sheep

Gray sheep đề cập đến những người dùng có ý kiến không nhất quán đồng ý hoặc không đồng ý với bất kỳ nhóm người nào do đó CF không có hiệu quả trong trường hợp này. *Black sheep* đề cập đến nhóm đối lập có thị hiếu đặc trưng đưa ra các khuyến nghị gần như không thể chẳng hạn như thích nhưng lại dùng những từ ngữ đánh giá như không thích do đó không thể gợi ý chính xác cho nhóm này.

1.4. Các kỹ thuật lọc cộng tác

Kỹ thuật lọc cộng tác được chia làm hai loại cơ bản là *Lọc cộng tác dựa trên bộ nhớ* và *Lọc cộng tác dựa trên mô hình*. Được thể hiện qua hình 1.2.



Hình 1.2: Các kỹ thuật lọc cộng tác

1.4.1. Kỹ thuật lọc cộng tác dựa trên bộ nhớ

Các hệ thống này sử dụng các kỹ thuật thống kê để xác định một nhóm người dùng được gọi là hàng xóm có lịch sử đồng ý với người dùng mục tiêu (tức là họ có đánh giá giống nhau về các sản phẩm khác nhau hoặc họ có xu hướng mua nhóm sản phẩm tương tự). Sau khi một vùng lân cận của người dùng được hình thành, các hệ thống này sử dụng các thuật toán khác nhau kết hợp các sở thích của những người hàng xóm để tạo ra một đề xuất dự đoán hoặc top-N sản phẩm cho người dùng đang hoạt động. Các kỹ thuật này còn được gọi là láng giềng gần nhất hoặc cộng tác dựa trên người dùng.

Kỹ thuật lọc cộng tác dựa trên bộ nhớ được chia làm 2 loại: *Lọc cộng tác dựa trên người dùng* và *Lọc cộng tác dựa trên sản phẩm*

1.4.1.1. Lọc cộng tác dựa trên người dùng

Đây là phương pháp sử dụng toàn bộ ma trận đánh giá để chọn ra một tập người dùng tương đồng nhất với người dùng cần được tư vấn. Sau đó, kết hợp các đánh giá của tập những người dùng tương đồng nhất này để đưa ra dự đoán cho người dùng cần được tư vấn về một sản phẩm chưa biết.

Các bước thực hiện tư vấn lọc cộng tác dựa trên người dùng:

Bước 1: Tiền xử lý dữ liệu: Dữ liệu được thu thập là những đánh giá về sản phẩm của người dùng. Dữ liệu thu thập thường rất lớn tuy nhiên trong đó một số đánh giá không có ích trong quá trình tư vấn theo phương pháp lọc cộng tác. Do đó cần tối ưu dữ liệu đầu vào bằng cách loại bỏ một số sản phẩm hoặc người dùng đánh giá quá ít sản phẩm, hoặc sản phẩm được quá ít người dùng đánh giá.

Bước 2: Tính toán mức độ tương tự của người dùng cần tư vấn với tất cả những người dùng trong hệ thống.

Bước 3: Xác định tập người dùng láng giềng với người dùng cần tư vấn bằng cách chọn K1 người dùng có mức độ tương tự với người dùng mục tiêu là cao nhất.

Bước 4: Dự đoán đánh giá của người dùng cần tư vấn với sản phẩm chưa đánh giá bằng việc kết hợp các đánh giá của những người dùng trong tập láng giềng.

Bước 5: Tư vấn K sản phẩm mới có mức độ phù hợp cao nhất cho người dùng cần tư vấn.

1.4.1.2. Lọc cộng tác dựa trên sản phẩm

Giải thuật lọc cộng tác dựa trên sản phẩm để tư vấn cho người dùng khác với giải thuật lọc cộng tác dựa trên người dùng bởi đối tượng được xét ở đây là các sản phẩm. Quá trình tư vấn bằng phương pháp lọc cộng tác dựa trên sản phẩm sẽ tính toán độ tương tự các sản phẩm, sau đó lựa chọn k sản phẩm tương tự $\{i_1, i_2,.., i_k\}$. Khi đó các dự đoán được tính toán dựa trên trung bình các đánh giá của người dùng trên những sản phẩm tương tự.

Các bước thực hiện tư vấn theo phương pháp lọc cộng tác dựa trên sản phẩm:

Bước 1: Tiền xử lý dữ liệu.

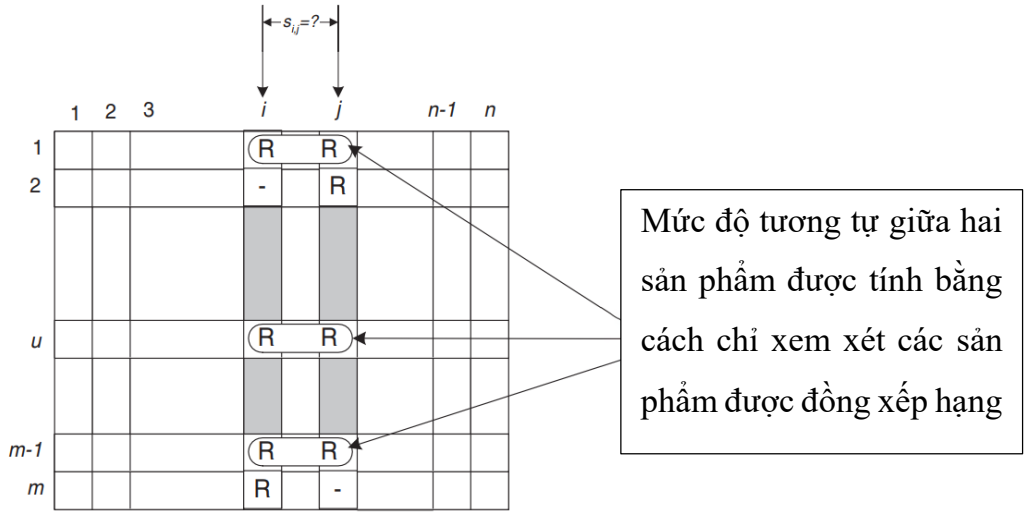
Bước 2: Xây dựng Ma trận đánh giá: Hàng là người dùng, Cột là các sản phẩm.

Bảng 1.2: Ma trận đánh giá

Người dùng	Sản phẩm						
	i_1	i_2	i_3	i_4	i_5	i_6	i_7
u_1	4			5	1		
u_2	5	5	4				
u_3				2	4	5	
u_4		3					3

Bước 3: Tính độ tương tự của các cặp sản phẩm, xây dựng *Ma trận tương tự* của các sản phẩm.

Một bước quan trọng trong thuật toán lọc cộng tác dựa trên sản phẩm là tính toán sự tương tự giữa các sản phẩm và sau đó chọn những sản phẩm có mức độ tương đồng cao nhất. Ý tưởng cơ bản tính toán độ tương tự giữa hai sản phẩm *i* và *j* là trước tiên tách những người dùng đã xếp hạng cả hai sản phẩm này và sau đó áp dụng một độ đo tương tự để tính toán mức độ tương tự $S_{i,j}$. Hình 1.3 minh họa quá trình này, ở đây các hàng của ma trận đại diện cho người dùng và các cột đại diện cho các sản phẩm. Có một số cách khác nhau để tính độ giống nhau giữa các sản phẩm sẽ được trình bày cụ thể trong chương 2.



Hình 1.3: Tách các sản phẩm cùng được đánh giá và tính toán độ tương tự

Trong trường hợp xét các sản phẩm i và j, sự giống nhau $S_{i,j}$ được tính bằng cách xem xét các sản phẩm đồng xếp hạng, tức là các cặp người dùng đã đánh giá cả 2 sản phẩm trên sau đó áp dụng các kỹ thuật tính toán độ tương tự để mô tả độ tương tự $S_{i,j}$, trong hình những người dùng cùng đánh giá sản phẩm i và j là 1,u và m-1.

Bước 4: Tính dự đoán của người dùng đối với sản phẩm dựa trên những sản phẩm lân cận với sản phẩm dự đoán được trình bày cụ thể trong mục 1.6.

Ví dụ minh họa thực tế về một hệ thống lọc cộng tác dựa trên sản phẩm: Giả sử sản phẩm ở đây là các bộ phim và người dùng là các khách hàng đăng nhập vào một hệ thống webstie để xem phim. Trên hệ thống sẽ lưu trữ hồ sơ của mỗi người dùng bao gồm thông tin cá nhân, các đánh giá của người dùng với các bộ phim. Các đánh giá được thực hiện theo thang điểm từ 1 sao đến 5 sao, đánh giá càng cao thì người dùng càng thích bộ phim đó. Công việc của hệ thống tư vấn là khi một người dùng đăng nhập vào hệ thống, hệ thống cần tư vấn những bộ phim cho người dùng đó và những bộ phim được tư vấn đó được dự đoán là người dùng sẽ đánh giá cao. Hệ thống xem xét các bộ phim mà người dùng chưa xem, so sánh độ tương tự giữa bộ phim đó với những bộ phim khác. Độ tương tự hai bộ phim được tính dựa trên những người dùng từng đánh giá trên cả hai bộ phim đó theo một thuật toán tính xác suất. Bước cuối cùng của hệ thống tư vấn là dự đoán đánh giá của người dùng với những bộ phim mà người dùng chưa xem, lựa chọn những bộ phim được dự đoán có đánh giá cao để đưa vào danh sách tư vấn cho người dùng.

1.4.2. Kỹ thuật lọc cộng tác dựa trên mô hình

1.4.2.1. Mô hình mạng Bayes

Mô hình mạng Bayes là một đồ thị có hướng, xoay chiều, trong đó mỗi nút n $\in$ N đại diện cho một biến ngẫu nhiên, mỗi cung có hướng a $\in$ A giữa các nút là một liên kết xác suất giữa các biến, và Θ là một bảng xác suất có điều kiện để định lượng mức độ phụ thuộc của một nút vào cha mẹ của nó. Mô hình mạng Bayer thường được sử dụng cho các nhiệm vụ phân loại.

Phương pháp mạng Bayes đơn giản cho lọc cộng tác được Breese đề xuất sử dụng một chiến lược Bayes (NB) gần nhất để đưa ra dự đoán cho các nhiệm vụ CF. Giả sử các tính năng là độc lập với lớp, xác suất của một lớp nhất định cho tất cả các đặc trưng có thể được tính toán, sau đó lớp có xác suất cao nhất sẽ được phân loại là lớp dự đoán. Đối với dữ liệu không đầy đủ, việc tính toán xác suất và sản xuất phân loại được tính trên dữ liệu quan sát. Breese giả thiết các giá trị đánh giá như những số nguyên ở giữa 0 và n, $r_{u,p}$ là đánh giá chưa biết của người dùng u đối với sản phẩm p được ước lượng dựa vào những đánh giá trước đó của người dùng u. Gọi P_u= {p' ∈ P| $r_{u,p'} \neq 0$}. Khi đó, đánh giá chưa biết của người dùng u đối với sản phẩm p được tính theo công thức (1.1).

$$r_{u,p} = E(r_{u,p}) = \sum_{1=0}^{n} i \times \Pr(r_{u,p} = i | r_{u,p'}, p' \in P_u) \quad (1.1)$$

Billsus và Pazzani đã chuyển đổi dữ liệu có nhiều mức đánh giá thành dữ liệu nhị phân. Khi đó ma trận đánh giá được chuyển đổi thành ma trận có các đặc trưng nhị phân. Việc chuyển đổi này làm cho việc sử dụng mô hình mạng Bayes trở nên thuận tiện hơn. Tuy nhiên, kết quả phân loại sử dụng các đặc trưng nhị phân không phản ánh đúng các bộ dữ liệu thực. Su và Khoshgoftaar [1] đã mở rộng mô hình mạng Bayes gồm nhiều lớp đánh giá khác nhau cho các tập dữ liệu thực. Mô hình đưa ra kết quả dự đoán tốt hơn so với các phương pháp dựa trên độ tương quan Pearson và mô hình mạng Bayes đơn giản.

1.4.2.2. Mô hình phân cụm

Một cụm là tập hợp các đối tượng dữ liệu tương tự với nhau và không giống với các đối tượng trong các cụm khác. Kỹ thuật phân cụm được thực hiện bằng cách phân chia các đối tượng dữ liệu ban đầu vào trong các cụm dữ liệu khác nhau nhằm mục đích tập trung khai thác thông tin từ đối tượng dữ liệu có quan hệ mật thiết với nhau, cũng như bỏ qua những thông tin nhiễu từ những đối tượng dữ liệu ít quan trọng.

Lọc cộng tác áp dụng các phương pháp phân cụm để phân chia tập người dùng (hoặc tập sản phẩm) thành các cụm người dùng (hoặc cụm sản phẩm) có sở thích tương tự nhau. Qua đó tư vấn các sản phẩm được đánh giá cao cho người dùng (hoặc sản phẩm) dựa vào cụm đó.

Phương pháp phân cụm có thể được phân thành ba loại: phương pháp phân vùng, phương pháp dựa trên mật độ và phương pháp phân cấp. Một phương pháp phân vùng thường được sử dụng là K-Means, do MacQueen đề xuất, có hai ưu điểm chính đó là hiệu quả tương đối và dễ thực hiện. Các phương pháp phân cụm dựa trên mật độ thường tìm kiếm các cụm đối tượng dày đặc được phân tách bằng các vùng thưa thớt biểu thị nhiễu. DBSCAN và OPTICS là các phương pháp phân cụm dựa trên mật độ của dữ liệu. Các phương pháp phân nhóm phân cấp, chẳng hạn như BIRCH tạo ra một phân tách phân cấp của tập hợp các đối tượng dữ liệu bằng cách sử dụng một số tiêu chí cụ thể.

Trong hầu hết các tình huống, phân cụm là một bước trung gian và các cụm kết quả được sử dụng để phân tích hoặc xử lý thêm để tiến hành phân loại hoặc thực hiện các nhiệm vụ khác. Sarwar và cộng sự, O'Connor và Herlocker đã sử dụng kỹ thuật phân cụm để phân vùng dữ liệu thành các cụm và sử dụng thuật toán lọc cộng tác dựa trên bộ nhớ chẳng hạn như thuật toán dựa trên tương quan Pearson để đưa ra dự đoán cho các tác vụ lọc cộng tác trong mỗi cụm.

1.5. Các tiêu chuẩn đánh giá độ đo

Chất lượng của một hệ thống giới thiệu có thể được quyết định dựa trên kết quả đánh giá. Loại số liệu được sử dụng phụ thuộc vào loại ứng dụng CF. Các số liệu đánh giá các hệ thống đề xuất có thể được phân loại rộng rãi thành các loại sau:

- Tiêu chuẩn đánh giá độ chính xác của đánh giá dự đoán bao gồm: Sai số tuyệt đối trung bình (MAE) và Sai số trung bình bình phương (RMSE)
- Tiêu chuẩn đánh giá độ chính xác của danh sách sản phẩm tư vấn bao gồm Độ chính xác (Precision), độ nhạy (Recall), E-measure, F-measure; Độ chính xác trung bình tuyệt đối (MAP)

1.5.1. Tiêu chuẩn đánh giá độ chính xác của đánh giá dự đoán

Sai số tuyệt đối trung bình (MAE)

Độ đo được sử dụng rộng rãi nhất trong lọc cộng tác để đánh giá sai số giữa giá trị đánh giá dự đoán và giá trị đánh giá thực tế là sai số tuyệt đối trung bình (MAE)

Sai số dự đoán MAE_u với mỗi người dùng u thuộc tập dữ liệu kiểm tra U_{test} được tính bằng trung bình cộng của sai số tuyệt đối giữa hai giá trị được dự đoán và xếp hạng thực của người dùng u với tất cả các sản phẩm thuộc tập P_u

$$MAE_u = \frac{1}{|P_u|}\sum_{x \in P_u} |\hat{r}_x^u - r_x^u| \quad (1.2)$$

Trong đó: P_u: là tổng xếp hạng trên tất cả người dùng

$\hat{r}_x^u$: là xếp hạng dự đoán của người dùng u trên sản phẩm x

r_x^u: là xếp hạng thực của người dùng u trên sản phẩm x

Sai số dự đoán trên tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số dự đoán cho mỗi người dùng thuộc U_{test}.

$$MAE = \frac{\sum_{u \in U_{test}} MAE_u}{|U_{test}|} \quad (1.3)$$

Sai số trung bình bình phương (RMSE)

Một tiêu chuẩn thông dụng khác cũng được dùng để đánh giá dự đoán là sai số trung bình bình phương RMSE. RMSE được tính bằng căn bậc hai của trung bình bình phương giữa xếp hạng thực và xếp hạng dự đoán.

$$RMSE_u = \sqrt{\frac{1}{|P_u|}\sum_{x \in P_u} (\hat{r}_x^u - r_x^u)^2} \quad (1.4)$$

$$RMSE_u = \frac{\sum_{u \in U_{test}} RMSE_u}{|U_{test}|} \quad (1.5)$$

Độ đo RMSE được sử dụng trong trường hợp chú trọng đặc biệt vào đánh giá độ chính xác cho những dự đoán có sai số lớn hơn so với giá trị thực tế.

Giá trị MAE, RMSE càng nhỏ thì hệ tư vấn cho kết quả càng chính xác.

1.5.2. Tiêu chuẩn đánh giá độ chính xác của danh sách sản phẩm tư vấn

Giá trị đánh giá của người dùng với sản phẩm được chia thành 2 loại: đánh giá "*thích*" và "*không thích*". Những đánh giá có giá trị lớn hơn hoặc bằng một ngưỡng cho trước được gọi là đánh giá "*thích*" và những đánh giá có giá trị nhỏ hơn ngưỡng giá trị cho trước gọi là đánh giá "*không thích*".

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn với những sản phẩm thực tế "*thích*" bởi người dùng, ta tiến hành xây dựng ma trận nhầm lẫn (Confusion matrix) sau:

Bảng 1.3: Ma trận nhầm lẫn

Đánh giá thực tế/Đánh giá dự đoán	Thích	Không thích
Thích	a	b
Không thích	c	d

Trong đó:

- Tổng số sản phẩm tư vấn cho các người dùng U_{test} là:

  - a là sản phẩm tư vấn cũng là sản phẩm thực tế thích bởi người dùng.

  - c là sản phẩm tư vấn nhưng lại là những sản phẩm thực tế không được thích bởi người dùng.

- Tổng số sản phẩm không được tư vấn cho những người dùng U_{test} là:

  - b là sản phẩm không được tư vấn nhưng thực tế sản phẩm này được thích bởi người dùng.

  - d là sản phẩm không được tư vấn và thực tế các sản phẩm này cũng không được thích bởi người dùng.

Từ ma trận nhầm lẫn, lĩnh vực học máy đã đưa ra một số độ đo tính chính xác của danh sách sản phẩm tư vấn như:

Độ chính xác (Precision), độ nhạy (Recall), E-measure, F-measure

Từ một tập các sản phẩm có sẵn của hệ thống, hệ tư vấn đưa ra một danh sách các sản phẩm người dùng có thể thích tương tự với quá trình lọc thông tin trong lĩnh vực truy vấn thông tin (Information Retrieval – IR). Do vậy, các độ đo tiêu chuẩn trong lọc thông tin thường được sử dụng để đánh giá hiệu năng của hệ tư vấn. Hai trong số các độ đo đó là độ chính xác (Precision), độ nhạy (Recall). Độ chính xác và độ nhạy được xác định theo công thức sau:

- Độ chính xác (P)

$$P = \frac{d}{b+d} \tag{1.6}$$

Trong đó:

- d sản phẩm tư vấn trong tập Q cũng là sản phẩm thực tế được thích bởi người dùng thuộc U_{test}.

- b + d: tổng số sản phẩm trong tập Q tư vấn cho người dùng U_{test}.

- Độ nhạy (R)

$$R = \frac{d}{c+d} \tag{1.7}$$

Trong đó:

- c + d: tổng số sản phẩm trong tập Q thực tế thích bởi những người dùng trong U_{test}.

Độ nhạy và độ chính xác có giá trị ngược nhau: Độ nhạy cao thì độ chính xác thấp và ngược lại. Để cân bằng giữa hai độ đo này, một độ đo mới được đưa ra đó là E-measure theo công thức sau:

$$E - measure = \frac{1}{\propto \frac{1}{P} + (1-\propto)\frac{1}{R}} \tag{1.8}$$

Tham số $\propto$ là độ lệch cho trước giữa P và R. Giá trị $\propto \in [0,1]$.

Trong trường hợp $\propto = 0.5$ khi đó P và R có vai trò như nhau trong việc đánh giá độ chính xác của hệ thống. Với trường hợp này, độ đo $E - \text{measure}$ được định nghĩa với tên mới $\text{F} - measure$ theo công thức sau:

$$\text{F} - \text{measure} = \frac{2\times\text{P}\times\text{R}}{\text{P}+\text{R}} = \frac{2}{\frac{1}{\text{P}}+\frac{1}{\text{R}}} \quad (1.9)$$

Giá trị $F - measure$ càng cao thì chứng tỏ hệ tư vấn cho kết quả càng chính xác.

Tuy nhiên trên thực tế thì số lượng kết quả trả về sẽ rất lớn mà người dùng không cần thiết lựa chọn hết nên chỉ số P không còn mấy ý nghĩa. Vì vậy người ta thường sử dụng chỉ số P@k để có thể đánh giá kết quả trả về chính xác hơn. P@k chỉ lựa chọn *top k* sản phẩm có giá trị dự đoán cao nhất để đánh giá.

$$\text{P@k} = \frac{|\text{P}_{\text{true}}|}{\min(\text{k},|\text{P}_\text{t}|)} \quad (1.10)$$

Trong đó:

- $|P_{true}|$: Tổng số kết quả chính xác trong *top k* sản phẩm được chọn tư vấn.
- $|P_t|$: Tổng số lượng kết quả.

$P@10$ thể hiện rằng cứ 10 gợi ý tới người dùng thì sẽ có bao nhiêu gợi ý được người dùng lựa chọn.

R@k cũng là một chỉ số đánh giá phổ biến thể hiện xác suất gợi ý thành công trong *top k* sản phẩm được chọn tư vấn, được tính bằng công thức sau:

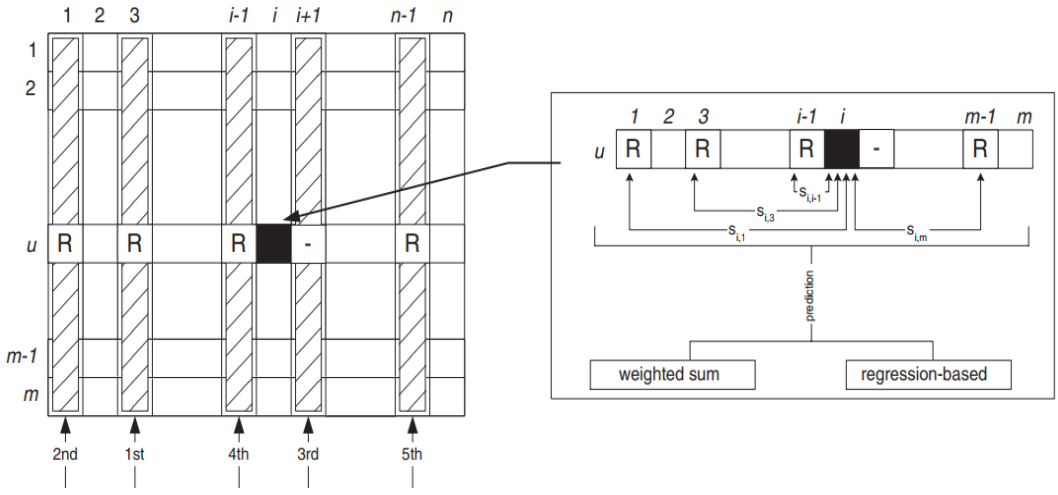
$$\text{R@}k = \frac{|P_{true}|}{\min(k,\text{M})} \quad (1.11)$$

Trong đó:

- $|P_{true}|$: Tổng số kết quả chính xác trong *top k* sản phẩm được chọn tư vấn.
- M: Tổng số lượng lựa chọn của người dùng.

1.6. Công thức dự đoán

Bước quan trọng nhất trong hệ thống lọc cộng tác là tạo giao diện đầu ra về mặt dự đoán. Sau khi tính toán mức độ tương tự giữa các sản phẩm thì bước tiếp theo là xem xét xếp hạng của người dùng mục tiêu và sử dụng một kỹ thuật để có được dự đoán.



Hình 1.4: Mô phỏng công thức dự đoán

1.6.1. Công thức dự đoán dựa trên người dùng

Để dự đoán xếp hạng của một người dùng đang hoạt động, nhiều biện pháp đã được đề xuất. Phép đo được sử dụng phổ biến nhất trong lĩnh vực này là phương pháp tổng có trọng số (Sarwar và cộng sự, 2001) được cho bởi công thức 1.12.

$$\tilde{r}_{u,i} = \frac{\sum_{v \in N_u^i} Sim_{uv} * r_{vi}}{\sum_{v \in N_u^i} |Sim_{uv}|} \quad (1.12)$$

Trong đó

N_u^i: Là tập những người hàng xóm giống nhất với người dùng u và đã xếp hạng mục i.

v: là người dùng thuộc N_u^i.

Sim_{uv}: là giá trị tương tự giữa người dùng u và v.

Ngoài ra chức năng dự đoán trung bình được đề xuất bởi (Aggarwal, 2016) là một biện pháp tổng hợp phổ biến được cho bởi công thức (1.13)

$$\tilde{r}_{u,i} = \overline{r_u} + \frac{\sum_{v \in N_u^i} Sim_{uv} * (r_{vi} - \bar{r}_v)}{\sum_{v \in N_u^i} |Sim_{uv}|} \quad (1.13)$$

Trong đó

$\overline{r_u}$: là xếp hạng trung bình của người dùng u.

$\bar{r}_v$: là xếp hạng trung bình của người dùng v.

1.6.2. Công thức dự đoán dựa trên sản phẩm

Tương tự như công thức dự đoán dựa trên người dùng, công thức dự đoán dựa theo sản phẩm được cho như công thức (1.14)

$$\tilde{r}_{u,i} = \frac{\sum_{j \in N_i^u} Sim_{ij} * r_{uj}}{\sum_{v \in N_i^u} |Sim_{ij}|} \quad (1.14)$$

Trong đó

N_i^u: là tập những người hàng xóm giống nhất với sản phẩm i và đã được người dùng u xếp hạng.

j: là người dùng thuộc N_i^u.

Sim_{ij}: là giá trị tương tự giữa sản phẩm i và j.

Công thức tổng hợp lấy giá trị trung bình được cung cấp trong (1.15)

$$\tilde{r}_{u,i} = \overline{r_i} + \frac{\sum_{j \in N_i^u} Sim_{ij} * (r_{uj} - \bar{r}_j)}{\sum_{v \in N_i^u} |Sim_{ij}|} \quad (1.15)$$

Trong đó

$\overline{r_i}$: là xếp hạng trung bình của sản phẩm i.

$\bar{r}_j$: là xếp hạng trung bình của sản phẩm j.

1.7. Kết luận

Trong chương này, luận văn đã trình bày về kỹ thuật lọc cộng tác bao gồm các kỹ thuật lọc cộng tác dựa trên bộ nhớ và lọc công tác dựa trên mô hình, các tiêu chuẩn đánh giá độ đo và các công thức dự đoán kết quả. Để dự đoán được kết quả trong tư vấn lọc cộng tác chúng ta phải sử dụng các độ đo để tính toán khoảng cách giữa người dùng hoặc sản phẩm. Trong chương tiếp theo, luận văn tìm hiểu về một số độ đo tương tự dùng trong kỹ thuật lọc cộng tác.

Chương 2. MỘT SỐ ĐỘ ĐO TƯƠNG TỰ CHO TƯ VẤN LỌC CỘNG TÁC

2.1. Giới thiệu chung

Việc tìm kiếm sự tương đồng giữa những người dùng là nhiệm vụ quan trọng nhất vì độ chính xác và chất lượng của các đề xuất chủ yếu dựa vào họ. Có nhiều thước đo độ tương đồng để tìm ra điểm giống nhau giữa người dùng và sản phẩm, điều này làm giảm mức độ gần gũi và mức độ tách biệt giữa người dùng và sản phẩm. Việc chọn một thước đo độ tương đồng hoàn hảo là rất quan trọng đối với lọc cộng tác và do đó đối với hệ thống tư vấn vì các thước đo độ tương đồng khác nhau sẽ cung cấp các kết quả khác nhau trong các bối cảnh thông tin khác nhau. Nói chung, tùy thuộc vào thuộc tính của người dùng hoặc thông tin sản phẩm hoặc bản thân số đo, độ đo tương tự cung cấp một giá trị số duy nhất là khoảng cách hoặc độ tương tự giữa hai người dùng hoặc sản phẩm [15]. Chương này sẽ tìm hiểu về một số độ đo tương tự cho tư vấn lọc cộng tác.

- Khoảng cách Euclide
- Chỉ số Jaccard
- Tương tự Cosine
- Hệ số tương quan Pearson
- Hệ số tương quan Pearson ràng buộc
- Tương quan Pearson dựa trên chức năng Sigmoid

2.2. Một số độ đo tương tự

2.2.1. Khoảng cách Euclide (Euclide distance)

Khoảng cách Euclide từ người dùng u đến người dùng v (hoặc từ sản phẩm i đến sản phẩm j) là độ dài của một đoạn thẳng giữa hai người dùng (hoặc sản phẩm) trong không gian Euclide. Trong điều kiện thực tế, mỗi người dùng được biểu diễn bằng tọa độ Descartes của nó đối với cơ sở của các sản phẩm (điều tương tự đối với

một sản phẩm được biểu thị đối với cơ sở của người dùng) và khoảng cách giữa hai người dùng (hoặc hai sản phẩm) là giá trị tuyệt đối hiệu số của tọa độ của chúng.

Công thức khoảng cách Euclide biểu thị mối tương quan giữa hai người dùng u và v như sau:

$$d_E(u,v) = \sqrt{\sum_{i \in I_{uv}} (r_{vi} - r_{ui})^2} \quad (2.1)$$

Trong đó

I_{uv}: là tập hợp các sản phẩm được xếp hạng bởi cả người dùng u và người dùng v.

r_{vi}: xếp hạng của người dùng v với sản phẩm i.

r_{ui}: xếp hạng của người dùng u với sản phẩm i.

Công thức khoảng cách Euclide biểu thị mối tương quan giữa hai sản phẩm i và j như sau:

$$d_E(i,j) = \sqrt{\sum_{u \in U_{ij}} \left(r_{uj} - r_{ui}\right)^2} \quad (2.2)$$

Trong đó

U_{ij}: là tập hợp những người dùng đã xếp hạng cả hai sản phẩm i và j.

r_{uj}: xếp hạng của người dùng u với sản phẩm j.

r_{ui}: xếp hạng của người dùng u với sản phẩm i.

Khoảng cách Euclide được chuẩn hóa thành độ đo tương tự Euclide cho người dùng và sản phẩm được thể hiện ở công thức (2.3) và (2.4).

$$Sim_E(u,v) = \frac{1}{1 + d_E(u,v)} \quad (2.3)$$

$$Sim_E(i,j) = \frac{1}{1 + d_E(i,j)} \quad (2.4)$$

2.2.2. Chỉ số Jaccard (Jaccard index)

Chỉ số Jaccard (Jaccard, 1912), ký hiệu là J tính độ giống nhau và đa dạng của hai tập hợp. Hệ số Jaccard giữa hai tập hữu hạn được định nghĩa là giao của hai tập hợp chia cho hợp của hai tập hợp. Có nghĩa là, nó đo tỷ lệ giữa số phần tử được chia sẻ giữa hai tập hợp với tổng số phần tử trong cả hai tập hợp. Chỉ số J nhận giá trị từ 0 đến 1, chỉ số càng gần 1 thì hai vectơ càng giống nhau.

Chỉ số Jaccard giữa hai người dùng u và v được tính theo công thức:

$$Sim_J(u,v) = \frac{|I_u| \cap |I_v|}{|I_u| \cup |I_v|} \quad (2.5)$$

Trong đó

I_u: Số lượng các sản phẩm do người dùng u đánh giá.

I_v: Số lượng các sản phẩm do người dùng v đánh giá.

Chỉ số Jaccard giữa hai sản phẩm i và j được tính theo công thức:

$$Sim_J(i,j) = \frac{|U_i| \cap |U_j|}{|U_i| \cup |U_j|} \quad (2.6)$$

Trong đó

U_i: Số lượng người dùng cùng đánh giá sản phẩm i.

U_j: Số lượng người dùng cùng đánh giá sản phẩm j.

2.2.3. Tương tự Cosine (Cosine similarity)

Kỹ thuật này được Breese và cộng sự trình bày bằng cách cho một người dùng ở dạng véc tơ xếp hạng do chính họ đánh giá và một sản phẩm dưới dạng véc tơ xếp hạng do nhóm người dùng đánh giá. Cosin giữa hai véc tơ đại diện cho hai người dùng hoặc hai sản phẩm cho biết giá trị tương tự giữa chúng. Giá trị gần bằng 1 cho biết nó tồn tài mối tương quan chặt chẽ giữa hai biến. Giá trị gần bằng 0 cho biết không có mối tương quan (các biến độc lập).

Độ tương tự cosin giữa hai người dùng u và v là cosin của 2 véc-tơ u và v được tính theo công thức (2.7). Trong đó, hai người dùng u và v được xem như hai véc-tơ m chiều, m =I_{uv} là số lượng các sản phẩm cả hai người dùng cùng đánh giá sản phẩm.

$$Sim_{COS}(u,v) = \cos(\vec{u},\vec{v}) = \frac{\vec{u}.\vec{v}}{\|\vec{u}\|^2 \times \|\vec{v}\|^2} = \frac{\sum_{i \in I_{uv}} r_{ui}.r_{vpi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2}\sqrt{\sum_{i \in I_v} r_{vi}^2}} \quad (2.7)$$

Độ tương tự cosin giữa hai sản phẩm i và j là cosin của 2 véc-tơ *i* và *j* được tính theo công thức (2.8). Trong đó, hai sản phẩm *i* và *j* được xem như hai véc-tơ n chiều, n =U_{ij} là số lượng người dùng cùng đánh giá sản phẩm sản phẩm i và j.

$$Sim_{Cos}(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}.\vec{j}}{\|\vec{i}\|^2 \times \|\vec{j}\|^2} = \frac{\sum_{u \in U_{ij}} r_{ui}.r_{uj}}{\sqrt{\sum_{u \in U_i} r_{ui}^2}\sqrt{\sum_{u \in U_j} r_{uj}^2}} \quad (2.8)$$

2.2.4. Hệ số tương quan Pearson (Pearson Correlation Coefficient)

Hệ số tương quan Pearson (PCC) là một trong những thước đo độ tương đồng truyền thống phổ biến và nổi bật nhất được đề xuất bởi Karl Pearson (Pearson, 1895) để đo các mối quan hệ tuyến tính và được sử dụng rộng rãi trong lĩnh vực thống kê, được biểu thị bằng tỷ lệ hiệp phương sai của hai người dùng hoặc sản phẩm với độ lệch chuẩn giữa người dùng hoặc sản phẩm, chỉ xem xét các sản phẩm được xếp hạng. Công thức PCC trả về giá trị từ -1 đến 1, trong đó: 1 cho biết mối tương quan dương mạnh mẽ, -1 cho thấy mối tương quan âm mạnh và 0 cho thấy không có mối tương quan nào cả (Resnick và cộng sự, 1994)

Hệ số tương quan Pearson giữa hai người dùng u, v được tính toán theo công thức (2.9)

$$Sim_{PCC}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2 \sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}} \quad (2.9)$$

Trong đó:

I_{uv} = $\{i \in I | r_{ui} \neq \emptyset \text{^} r_{vi} \neq \emptyset\}$ là tập hợp tất cả những sản phẩm người dùng u và v cùng đánh giá.

r_{ui}: Đánh giá của người dùng u cho sản phẩm i

r_{vi}: Đánh giá của người dùng v cho sản phẩm i

$\bar{r}_u$: là trung bình cộng các đánh giá khác ∅ của người dùng u

$\bar{r}_v$ là trung bình cộng các đánh giá khác ∅ của người dùng v

Hệ tương quan Pearson giữa hai sản phẩm i và j [20] được tính toán theo công thức (2.10)

$$Sim_{PCC}(i,j) = \frac{\sum_{u \in U_{ij}}(r_{ui}-\bar{r}_i)(r_{uj}-\bar{r}_j)}{\sqrt{\sum_{u \in U_{ij}}(r_{ui}-\bar{r}_i)^2 \sum_{u \in U_{ij}}(r_{uj}-\bar{r}_j)^2}} \quad (2.10)$$

Trong đó:

$U_{ij} = \{u \in U | r_{ui} \neq \emptyset \text{^} r_{uj} \neq \emptyset\}$ là tập hợp tất cả những người dùng cùng đánh giá sản phẩm i và j.

$\bar{r}_i$: là đánh giá trung bình cho sản phẩm i

$\bar{r}_j$ là đánh giá trung bình cho sản phẩm j.

2.2.5. Hệ số tương quan Pearson ràng buộc (Constrained Pearson Correlation)

Đề xuất RINGO được phát triển để cung cấp cho người dùng các đề xuất về album nhạc và nghệ sĩ. Trong RINGO, người dùng cung cấp phản hồi trên thang điểm danh nghĩa từ một ("không thích mạnh") đến bảy ("thích mạnh"), với giá trị trung bình tính ở giữa thang. Dựa trên số lượng ngày càng tăng của người dùng RINGO, Shraddhanand và Mae [5] đã đề xuất phương pháp tương quan Pearson có giới hạn (CPCC) để thay thế các biến xếp hạng trung bình được sử dụng bởi các phương pháp tiếp cận PCC bằng giá trị trung bình của thang đo xếp hạng tích cực và tiêu cực. Mối tương quan giữa hai người dùng *u* và *v* được tính như sau:

$$Sim_{CPCC}(u,v) = \frac{\sum_{i \in I_{uv}}(r_{ui}-r_m)(r_{vi}-r_m)}{\sqrt{\sum_{i \in I_{uv}}(r_{ui}-r_m)^2 \sum_{i \in I_{uv}}(r_{vi}-r_m)^2}} \quad (2.11)$$

Trong đó:

r_m: biểu thị giá trị trung bình của thang đánh giá. Ví dụ: trong thang điểm từ 1 đến 5 thì r_m là 3, r_m là 4 trong thang điểm từ 1 đến 7.

Mối tương quan giữa hai sản phẩm *i* và *j* được tính như sau:

$$Sim_{CPCC}(i,j) = \frac{\sum_{i \in U_{ij}}(r_{ui}-r_m)(r_{uj}-r_m)}{\sqrt{\sum_{i \in U_{ij}}(r_{ui}-r_m)^2 \sum_{i \in U_{ij}}(r_{uj}-r_m)^2}} \quad (2.12)$$

Hạn chế chính của kỹ thuật này là hiệu suất kém đối với tập dữ liệu thưa thớt.

2.2.6. Tương quan Pearson dựa trên chức năng Sigmoid (Sigmoid Function-Based Pearson Correlation)

Một phương pháp dựa trên tương quan Pearson khác là Hệ số tương quan Pearson dựa trên hàm Sigmoid (SPCC). Biện pháp tương tự này hoạt động tốt nhất khi có một số lượng lớn các sản phẩm được đồng xếp hạng giữa những người dùng. Mức độ tương tự giữa hai người dùng u và v trong SPCC được tính như sau:

$$Sim_{SPCC}(u,v) = Sim_{PCC}(u,v) \frac{1}{1+\exp(-\frac{|i'|}{2})} \quad (2.13)$$

Trong đó i' là tổng số người dùng đồng xếp hạng.

Mức độ tương tự giữa hai sản phẩm *i* và *j* trong SPCC được tính như sau:

$$Sim_{SPCC}(u,v) = Sim_{PCC}(i,j) \frac{1}{1+\exp(-\frac{|j'|}{2})} \quad (2.14)$$

Trong đó j' là tổng số sản phẩm đồng xếp hạng.

2.3. Ví dụ

Cho ma trận đánh giá R = (r_{ij}) trong hệ tư vấn lọc cộng tác gồm 7 người dùng U = {u_1, u_2, u_3, u_4, u_5, u_6, u_7} và 5 sản phẩm I = {i_1, i_2, i_3, i_4, i_5}. Mỗi người dùng đều đưa ra các đánh giá về các sản phẩm theo thang bậc {∅, 1, 2, 3, 4, 5}. Người dùng u_i chưa đánh giá sản phẩm i_j thì giá trị r_{ij}=∅ được thể hiện trong Bảng 2.1.

Bảng 2.1: Ma trận đánh giá của người dùng

	i_1	i_2	i_3	i_4	i_5
u_1	3	4		2	2
u_2	5		4	2	1
u_3	2		1		5
u_4	3	2		4	
u_5	1			5	
u_6		2	1		
u_7			1	4	4

Yêu cầu đặt ra của bài toán là tính toán mức độ tương tự giữa các cặp người dùng và các cặp sản phẩm. Vấn đề sẽ được giải quyết trong mục 2.3.1 và 2.3.2.

2.3.1. Độ tương tự giữa các cặp người dùng

Phần này trình bày cách tính độ tương tự giữa các cặp người dùng dựa vào các công thức tính độ đo tương tự: Khoảng cách Euclide, Chỉ số Jaccard, Tương tự Cosine, Hệ số tương quan Pearson, Hệ số tương quan Pearson ràng buộc, Hệ số tương quan Pearson dựa trên chức năng Sigmoid.

Khoảng cách Euclide biểu thị mối tương quan giữa hai người dùng u_1 và u_2 được tính như sau:

$$d_E(u_1, u_2) = \sqrt{(3-5)^2 + (2-2)^2 + (2-1)^2} = 2.24$$

Độ đo tương tự Euclide được tính như sau:

$$Sim_E(u_1, u_2) = \frac{1}{1+d_E(u_1,u_2)} = \frac{1}{1+2.24} = 0.31$$

Tương tự ta có bảng độ đo tương tự giữa hai người dùng sử dụng độ đo tương tự Euclide.

Bảng 2.2: Bảng tính độ tương tự giữa hai người dùng theo công thức công thức E

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.31	0.24	0.26	0.22	0.33	0.26
u_2	0.31	1.00	0.15	0.26	0.17	0.25	0.18
u_3	0.24	0.15	1.00	0.50	0.50	1.00	0.50
u_4	0.26	0.26	0.50	1.00	0.31	1.00	1.00
u_5	0.22	0.17	0.50	0.31	1.00	1.00	0.50
u_6	0.33	0.25	1.00	1.00	1.00	1.00	1.00
u_7	0.26	0.18	0.50	1.00	0.50	1.00	1.00

Chỉ số Jaccard giữa hai người dùng u_1 và u_2 được tính như sau:

$$Sim_J(u_1, u_2) = \frac{3}{5} = 0.60$$

Tương tự như vậy ta có bảng tính độ tương tự giữa hai người dùng khác nhau bằng chỉ số Jaccard

Bảng 2.3: Bảng tính độ tương tự giữa hai người dùng theo công thức J

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.60	0.40	0.60	0.50	0.20	0.40
u_2	0.60	1.00	0.75	0.40	0.50	0.20	0.75
u_3	0.40	0.75	1.00	0.20	0.25	0.25	0.50
u_4	0.60	0.40	0.20	1.00	0.67	0.25	0.20
u_5	0.50	0.50	0.25	0.67	1.00	0.00	0.25
u_6	0.20	0.20	0.25	0.25	0.00	1.00	0.25
u_7	0.40	0.75	0.50	0.20	0.25	0.25	1.00

Tương tự Cosine

Từ ma trận đánh giá của 7 người dùng về 5 sản phẩm ta tính giá trị trung bình cộng các đánh giá của người dùng.

Bảng 2.4: Giá trị trung bình cộng các đánh giá của người dùng

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
$\overline{r_{u_J}}$	2.75	3.00	2.67	3.00	3.00	1.50	3.00

Từ đó ta đưa ra ma trận chuẩn hóa dữ liệu

Bảng 2.5: Ma trận chuẩn hóa dữ liệu

	i_1	i_2	i_3	i_4	i_5
u_1	0.25	1.25		-0.75	-0.75
u_2	2.00		1.00	-1.00	-2.00
u_3	-0.67		-1.67		2.33
u_4	0.00	-1.00		1.00	
u_5	-2.00			2.00	
u_6		0.50	-0.50		
u_7			-2.00	1.00	1.00

Độ tương tự cosin giữa hai người dùng u_1, u_2 được tính như sau:

$$Sim_{COS}(u_1, u_2) = \frac{0.25\times2+(-0.75)\times(-1)+(-0.75)\times(-2)}{\sqrt{0.25^2+1.25^2+(-0.75)^2+(-0.75)^2}\times\sqrt{2^2+1^2+(-1)^2+(-2)^2}}$$

$$Sim_{COS}(u_1, u_2) = 0.52$$

Tương tự như vậy ta có bảng tính độ tương tự giữa hai người dùng khác nhau bằng độ tương tự cosin (Bảng 2.6).

Bảng 2.6: Bảng tính độ tương tự giữa hai người dùng theo công thức COS

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.52	-0.39	-0.85	-0.43	0.53	-0.37
u_2	0.52	1.00	-0.82	-0.22	-0.67	-0.22	-0.65
u_3	-0.39	-0.82	1.00	0.00	0.16	0.40	0.79
u_4	-0.85	-0.22	0.00	1.00	0.50	-0.50	0.29
u_5	-0.43	-0.67	0.16	0.50	1.00	0.00	0.29
u_6	0.53	-0.22	0.40	-0.50	0.00	1.00	0.58
u_7	-0.37	-0.65	0.79	0.29	0.29	0.58	1.00

Hệ số tương quan Pearson giữa hai người dùng $u_1$và u_2 được tính như sau:

$$Sim_{PCC}(u_1, u_2) = \frac{(3-2.75)(5-3)+(2-2.75)(2-3)+(2-2.75)(1-3)}{\sqrt{(3-2.75)^2+(2-2.75)^2+(2-2.75)^2}\sqrt{(5-3)^2+(2-3)^2+(1-3)^2}}$$

$Sim_{PCC}(u_1, u_2)$= 0.84

Tương tự như vậy ta có bảng tính độ tương tự giữa hai người dùng theo công thức của hệ số tương quan

Bảng 2.7: Bảng tính độ tương tự giữa hai người dùng theo công thức PCC

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.84	-0.72	-0.96	-0.89	0.71	-1.00
u_2	0.84	1.00	-0.82	-0.45	-0.95	-1.00	-0.83
u_3	-1.00	-0.87	1.00	0.00	1.00	1.00	0.88
u_4	-0.96	-0.45	0.00	1.00	0.00	-1.00	1.00
u_5	-0.89	-0.95	1.00	0.00	1.00	0.00	1.00
u_6	0.71	-1.00	1.00	-1.00	0.00	1.00	1.00
u_7	-1.00	-0.83	0.88	1.00	1.00	1.00	1.00

Hệ số tương quan Pearson ràng buộc

Nhìn vào ma trận đánh giá người dùng (Bảng 2.1) ta thấy các đánh giá của người dùng từ 1 đến 5 từ đó ta có $r_m = 3$

Độ tương tự giữa hai người dùng u_1 và u_2 được tính như sau:

$$Sim_{CPCC}(u_1, u_2) = \frac{(3-3)\times(5-3)+(2-3)\times(2-3)+(2-3)\times(1-3)}{\sqrt{((3-3)^2+(2-3)^2+(2-3)^2)\times((5-3)^2+(2-3)^2+(1-3)^2)}}$$

$$Sim_{CPCC}(u_1, u_2) = 0.71$$

Tương tự ta có bảng tính toán độ tương tự giữa hai người dùng khác nhau dựa trên Hệ số tương quan Pearson ràng buộc.

Bảng 2.8: Bảng tính độ tương tự giữa hai người dùng theo công thức CPCC

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.71	-0.89	-1.00	-0.71	-1.00	-1.00
u_2	0.71	1.00	-0.89	-0.45	-0.95	-1.00	-0.83
u_3	-0.89	-0.89	1.00	0.00	1.00	1.00	0.95
u_4	-1.00	-0.45	0.00	1.00	0.00	1.00	1.00
u_5	-0.71	-0.95	1.00	0.00	1.00	0.00	1.00
u_6	-1.00	-1.00	1.00	1.00	0.00	1.00	1.00
u_7	-1.00	-0.83	0.95	1.00	1.00	1.00	1.00

Hệ số tương quan Pearson dựa trên chức năng Sigmoid

Độ tương tự giữa hai người dùng u_1 và u_2 được tính như sau:

$$Sim_{SPCC}(u_1, u_2) = 0.84 \times \frac{1}{1+\exp(-\frac{|3|}{2})} = 0.69$$

Tương tự ta có độ tương tự giữa hai người dùng khác nhau được thể hiện trong bảng 2.9.

Bảng 2.9: Bảng tính độ tương tự giữa hai người dùng theo công thức SPCC

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	1.00	0.69	-0.53	-0.78	-0.65	0.44	-0.73
u_2	0.69	1.00	-0.67	-0.33	-0.69	-0.62	-0.68
u_3	-0.53	-0.67	1.00	0.00	0.62	0.62	0.65
u_4	-0.78	-0.33	0.00	1.00	0.00	-0.62	0.62
u_5	-0.65	-0.69	0.62	0.00	1.00	0.00	0.62
u_6	0.44	-0.62	0.62	-0.62	0.00	1.00	0.62
u_7	-0.73	-0.68	0.65	0.62	0.62	0.62	1.00

Từ các bảng trên ta có bảng tổng hợp kết quả tính toán độ tương tự giữa các cặp người dùng dựa trên các công thức tính toán độ tương tự như sau:

Bảng 2.10: Bảng tổng hợp tính độ tương tự giữa hai người dùng

	E	**J**	**COS**	**PCC**	**CPCC**	**SPCC**
u_{12}	**0.31**	0.60	0.52	0.84	0.71	0.69
u_{13}	0.24	0.40	-0.39	**-1.00**	**-0.89**	-0.53
u_{14}	0.26	0.60	-0.85	**-0.96**	**-1.00**	-0.78
u_{15}	0.22	0.50	-0.43	**-0.89**	**-0.71**	-0.65
u_{16}	0.33	0.20	0.53	**0.71**	**-1.00**	0.44
u_{17}	**0.26**	**0.40**	**-0.37**	-1.00	-1.00	-0.73
u_{23}	0.15	**0.75**	-0.82	**-0.87**	-0.89	-0.67
u_{24}	0.26	0.40	-0.22	-0.45	-0.45	-0.33
u_{25}	0.17	0.50	-0.67	**-0.95**	**-0.95**	-0.69
u_{26}	0.25	0.20	-0.22	**-1.00**	**-1.00**	-0.62
u_{27}	0.18	**0.75**	-0.65	**-0.83**	**-0.83**	**-0.68**
u_{34}	**0.50**	0.20	0.00	0.00	0.00	0.00

	E	J	COS	PCC	CPCC	SPCC
u_{35}	0.50	0.25	0.16	**1.00**	**1.00**	0.62
u_{36}	**1.00**	0.25	0.40	**1.00**	**1.00**	0.62
u_{37}	0.50	0.50	0.79	0.88	0.95	0.65
u_{45}	0.31	0.67	0.50	**0.00**	**0.00**	**0.00**
u_{46}	**1.00**	0.25	-0.50	**-1.00**	**1.00**	-0.62
u_{47}	**1.00**	0.20	0.29	**1.00**	**1.00**	0.62
u_{56}	**1.00**	0.00	0.00	0.00	0.00	0.00
u_{57}	0.50	0.25	0.29	**1.00**	**1.00**	0.62
u_{67}	1.00	0.25	0.58	**1.00**	**1.00**	0.62

Từ bảng 2.10 và bảng 2.1 ta đưa ra một số so sánh, đánh giá các độ đo tương tự với các cặp người dùng như sau:

- **Khoảng cách Euclide (E)**: Đối với cặp người dùng u_{12} trong bảng 2.1 cùng đánh giá 3/5 sản phẩm chiếm 60% tổng số sản phẩm, trong đó có 2/3 sản phẩm được đánh giá với mức xếp hạng tương đương nhau chiếm 67%. Như vậy độ tương tự giữa hai người dùng này ở mức trung bình đến tương đối cao. Tuy nhiên, trong bảng 2.10 độ tương tự giữa hai người dùng u_{12} được thể hiện ở mức độ thấp.

Đối với cặp u_{17}, số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, xếp hạng giống nhau trên cả 2 sản phẩm chiếm 100% như vậy độ tương tự giữa cặp người dùng u_{17} tương đối cao, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá thấp (bằng 0.26).

Đối với cặp u_{34}, số lượng sản phẩm cùng đánh giá là 1/5 chiếm 20% trên tổng số sản phẩm, mức xếp hạng khác nhau, như vậy độ tương tự giữa cặp người dùng u_{34} tương đối thấp, trong bảng 2.10 mức độ tương tự được đánh giá ở mức trung bình (bằng 0.5).

Đối với các cặp u_{36}, u_{46}, u_{47}, u_{67} số lượng sản phẩm cùng đánh giá là 1/5 chiếm 20% trên tổng số sản phẩm với mức xếp hạng tương tự nhau ở cả 2 sản phẩm chiếm 100%, như vậy độ tương tự giữa những cặp người dùng này chưa thể đánh giá ở mức

độ chính xác do số lượng sản phẩm cùng đánh giá ở mức thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá ở mức rất cao (bằng 1).

Đối với cặp u_{56}, số lượng sản phẩm cùng đánh giá là 0/5 chiếm 0% trên tổng số sản phẩm, như vậy độ tương tự giữa cặp người dùng này là 0. Tuy nhiên trong bảng 2.10 mức độ tượng tự được đánh giá rất cao (bằng 1).

Như vậy trong một số trường hợp *Khoảng cách Euclide (E)* đưa ra mức độ tương tự chưa chính xác.

- **Chỉ số Jaccard (J)**: Đối với cặp u_{17}, số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, xếp hạng giống nhau trên cả 2 sản phẩm chiếm 100% như vậy độ tương tự giữa cặp người dùng u_{17} tương đối cao, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá thấp (bằng 0.40).

Đối với cặp u_{23}, số lượng sản phẩm cùng đánh giá là 3/5 chiếm 67% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 3 sản phẩm, như vậy độ tương tự giữa cặp người dùng u_{23} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá khá cao (bằng 0.75).

Tương tự với cặp u_{27}, số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 2 sản phẩm, như vậy độ tương tự giữa cặp người dùng u_{23} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá khá cao (bằng 0.75).

- **Tương tự Cosin (COS)**: Đối với cặp u_{17}, số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, xếp hạng giống nhau trên cả 2 sản phẩm chiếm 100% như vậy độ tương tự giữa cặp người dùng u_{17} tương đối cao, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá thấp (bằng 0.37)

- **Hệ tương quan Pearson (PCC) và Hệ tương quan Pearson ràng buộc (CPCC)**: Trong bảng 2.10 ta thấy mức độ tương tự giữa các cặp người dùng sử dụng công thức PCC và CPCC là tương đối giống nhau.

Đối với các cặp u_{13}, u_{25} số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 2 sản phẩm, như vậy độ tương tự giữa các cặp người dùng u_{13}, u_{25} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá cao.

Đối với các cặp u_{14}, u_{23}, u_{27} số lượng sản phẩm cùng đánh giá là 3/5 chiếm 67% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 3 sản phẩm, như vậy độ tương tự giữa các cặp người dùng u_{13}, u_{25}, u_{27} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá cao.

Đối với các cặp u_{16}, u_{26} số lượng sản phẩm cùng đánh giá là 1/5 chiếm 20% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 3 sản phẩm, như vậy độ tương tự giữa các cặp người dùng u_{13}, u_{25} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá cao.

Đối với các cặp u_{45} số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, mức độ xếp hạng tương tự nhau 1/2 sản phẩm chiếm 50%, như độ tương tự giữa các cặp người dùng u_{45} trung bình, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá là không có (bằng 0).

Đối với các cặp u_{35}, u_{36}, u_{46}, u_{47}, u_{57}, u_{67} số lượng sản phẩm cùng đánh giá là 1/5 chiếm 20% trên tổng số sản phẩm với mức xếp hạng tương tự nhau ở cả 2 sản phẩm chiếm 100%, như vậy độ tương tự giữa những cặp người dùng này chưa thể đánh giá ở mức độ chính xác cao hoặc thấp do số lượng sản phẩm cùng đánh giá ở mức thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá ở mức rất cao (bằng 1).

Như vậy hạn chế của PCC và CPCC là chỉ tính đến mức đánh giá tương tự của người dùng mà chưa tính đến số lượng các cặp sản phẩm cùng được đánh giá.

- **Tương quan Pearson dựa trên chức năng Sigmoid (SPCC)**:

Đối với các cặp u_{27} số lượng sản phẩm cùng đánh giá là 3/5 chiếm 67% trên tổng số sản phẩm, mức độ xếp hạng hoàn toàn khác nhau trên cả 3 sản phẩm, như vậy

độ tương tự giữa các cặp người dùng u_{13}, u_{25} thấp, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá cao.

Đối với các cặp u_{45} số lượng sản phẩm cùng đánh giá là 2/5 chiếm 40% trên tổng số sản phẩm, mức độ xếp hạng tương tự nhau 1/2 sản phẩm chiếm 50%, như độ tương tự giữa các cặp người dùng u_{45} trung bình, tuy nhiên trong bảng 2.10 mức độ tương tự được đánh giá là không có (bằng 0).

Như vậy nhìn vào kết quả tính toán độ tương tự với các độ đo khác nhau trên bảng 2.10 ta thấy độ tương tự giữa hai người dùng sử dụng các độ đo khác nhau sẽ cho kết quả khác nhau, tuy nhiên hai độ đo PCC và CPCC cho kết quả tương tự nhau và bị hạn chế nhiều so với các độ đo khác.

2.3.2. Độ tương tự giữa các cặp sản phẩm

Phần này trình bày cách tính độ tương tự giữa các cặp sản phẩm dựa vào các công thức tính độ đo tương tự: Khoảng cách Euclide, Chỉ số Jaccard, Tương tự Cosine, Hệ số tương quan Pearson, Hệ số tương quan Pearson ràng buộc, Hệ số tương quan Pearson dựa trên chức năng Sigmoid.

Khoảng cách Euclide biểu thị mối tương quan giữa hai sản phẩm i_1 và i_2 được tính như sau:

$$d_E(i_1, i_2) = \sqrt{(3-4)^2 + (3-2)^2} = 1.41$$

Độ đo tương tự Euclide được tính như sau:

$$Sim_E(i_1, i_2) = \frac{1}{1+d_E(i_1,i_2)} = \frac{1}{1+1.41} = 0.41$$

Tương tự ta có bảng độ đo tương tự giữa hai người dùng sử dụng độ đo tương tự Euclide (Bảng 2.11).

Bảng 2.11: Bảng tính độ tương tự giữa hai sản phẩm theo công thức E

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.41	0.41	0.16	0.16
i_2	0.41	1.00	0.50	0.26	0.33
i_3	0.41	0.50	1.00	0.22	0.15
i_4	0.16	0.26	0.22	1.00	0.50
i_5	0.16	0.33	0.15	0.50	1.00

Chỉ số Jaccard (Jaccard index)

Chỉ số Jaccard giữa hai sản phẩm i_1 và i_2 được tính như sau:

$$Sim_J(i_1, i_2) = \frac{2}{6} = 0.33$$

Tương tự như vậy ta có bảng tính độ tương tự giữa hai sản phẩm khác nhau bằng chỉ số Jaccard

Bảng 2.12: Bảng tính độ tương tự giữa hai sản phẩm theo công thức J

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.33	0.29	0.57	0.43
i_2	0.33	1.00	0.17	0.33	0.17
i_3	0.29	0.17	1.00	0.29	0.60
i_4	0.57	0.33	0.29	1.00	0.50
i_5	0.43	0.17	0.60	0.50	1.00

Tương tự Cosine (Cosine similarity)

Từ ma trận đánh giá của người dùng u_1, u_2, u_3, u_4 u_5, u_6, u_7 về các sản phẩm i_1, i_2, i_3, i_4, i_5 ta tính giá trị trung bình cộng các đánh giá của người dùng đối với từng sản phẩm như trong bảng 2.13.

Bảng 2.13: Giá trị trung bình cộng đánh giá từng sản phẩm

	i_1	i_2	i_3	i_4	i_5
$\overline{r_{i_j}}$	2.80	2.67	1.75	3.40	3.00

Từ đó ta có ma trận chuẩn hóa dữ liệu

Bảng 2.14: Ma trận chuẩn hóa dữ liệu

	i_1	i_2	i_3	i_4	i_5
u_1	0.20	1.33		-1.40	-1.00
u_2	2.20		2.25	-1.40	-2.00
u_3	-0.80		-0.75		2.00
u_4	0.20	-0.67		0.60	
u_5	-1.80			1.60	
u_6		-0.67	-0.75		
u_7			-0.75	0.60	1.00

Độ tương tự cosine giữa hai sản phẩm i_1 và i_2 được tính như sau:

$$Sim_{COS}(i_1, i_2) = \frac{0.2 \times 1.33 + 0.2 \times (-0.67)}{\sqrt{0.2^2 + 2.2^2 + (-0.8)^2 + 0.2^2 + (-1.8)^2}\sqrt{(1.33)^2 + (-0.67)^2 + (-0.67)^2}}$$

$$Sim_{COS}(i_1, i_2) = 0.03$$

Tương tự như vậy ta có bảng tính độ tương tự giữa hai sản phẩm khác nhau bằng độ tương tự cosine (Bảng 2.15).

Bảng 2.15: Bảng tính độ tương tự giữa hai sản phẩm theo công thức COS

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.03	0.72	-0.77	-0.66
i_2	0.03	1.00	0.12	-0.52	-0.26
i_3	0.72	0.12	1.00	-0.52	-0.82
i_4	-0.77	-0.52	-0.52	1.00	0.57
i_5	-0.66	-0.26	-0.82	0.57	1.00

Hệ tương quan Pearson giữa hai sản phẩm i_1 và i_2 được tính như sau:

$$Sim_{PCC}(i_1, i_2) = \frac{(3-2.8)(4-2.67)+(3-2.8)(2-2.67)}{\sqrt{(3-2.8)^2+(3-2.8)^2}\sqrt{(4-2.67)^2+(2-2.67)^2}}$$

$$Sim_{PCC}(i_1, i_2) = 0.32$$

Tương tự như vậy ta có bảng tính độ tương tự giữa hai sản phẩm theo công thức của hệ số tương quan (bảng 2.13)

Bảng 2.16: Bảng tính độ tương tự giữa hai sản phẩm theo công thức PCC

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.32	1.00	-0.82	-0.88
i_2	0.32	1.00	1.00	-1.00	-1.00
i_3	1.00	1.00	1.00	-1.00	-0.90
i_4	-0.82	-1.00	-1.00	1.00	0.95
i_5	-0.88	-1.00	-0.90	0.95	1.00

Hệ số tương quan Pearson ràng buộc:

Độ tương tự giữa hai người dùng i_1 và i_2 được tính như sau:

$$Sim_{CPCC}(i_1, i_2) = \frac{(3-3)\times(4-3)+(3-3)\times(2-3)}{\sqrt{((3-3)^2+(3-3)^2)\times((4-3)^2+(2-3)^2)}} = 0$$

Tương tự ta có bảng tính toán độ tương tự giữa hai sản phẩm khác nhau dựa trên Hệ số tương quan Pearson ràng buộc.

Bảng 2.17: Bảng tính độ tương tự giữa hai sản phẩm theo công thức CPCC

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.00	0.80	-0.80	-0.89
i_2	0.00	1.00	1.00	-1.00	-1.00
i_3	0.80	1.00	1.00	-0.95	-0.89
i_4	-0.80	-1.00	-0.95	1.00	0.94
i_5	-0.89	-1.00	-0.89	0.94	1.00

Tương quan Pearson dựa trên chức năng Sigmoid

Độ tương tự giữa hai sản phẩm i_1 và i_2 được tính như sau:

$$Sim_{SPCC}(i_1, i_2) = 0.32 \times \frac{1}{1+\exp(-\frac{|2|}{2})} = 0.23$$

Tương tự ta có độ tương tự giữa hai sản phẩm khác nhau được thể hiện trong bảng

Bảng 2.18: Bảng tính độ tương tự giữa hai sản phẩm theo công thức SPCC

	i_1	i_2	i_3	i_4	i_5
i_1	1.00	0.23	0.73	-0.72	-0.72
i_2	0.23	1.00	0.62	-0.73	-0.62
i_3	0.73	0.62	1.00	-0.73	-0.74
i_4	-0.72	-0.73	-0.73	1.00	0.77
i_5	-0.72	-0.62	-0.74	0.77	1.00

Từ các bảng trên ta có bảng tổng hợp kết quả tính toán độ tương tự giữa các cặp sản phẩm dựa trên các công thức tính toán độ tương tự như sau:

Bảng 2.19: Bảng tổng hợp tính độ tương tự giữa hai sản phẩm

	E	J	COS	PCC	CPCC	SPCC
i_{12}	0.41	0.33	0.03	0.32	0.00	0.23
i_{13}	0.41	0.29	0.72	**1.00**	0.80	0.73
i_{14}	0.16	0.57	-0.77	-0.82	-0.80	-0.72
i_{15}	0.16	0.43	-0.66	-0.88	-0.89	-0.72
i_{23}	0.50	0.17	0.12	**1.00**	**1.00**	0.62
i_{24}	0.26	0.33	-0.52	**-1.00**	**-1.00**	-0.73
i_{25}	0.33	0.17	-0.26	**-1.00**	**-1.00**	-0.62
i_{34}	0.22	0.29	-0.52	**-1.00**	-0.95	-0.73
i_{35}	0.15	0.60	-0.82	-0.90	-0.89	-0.74
i_{45}	0.50	0.50	0.57	0.95	0.94	0.77

Từ bảng 2.19 ta thấy một số hạn chế của hai độ đo tương tự PCC và CPCC như sau:

- **Hệ tương quan Pearson (PCC)**: Đối với các cặp sản phẩm i_{12} và i_{23} đều có chung số lượng người dùng đánh giá là 1/7 chiếm 14% trong tổng số người dùng với mức đánh giá tương đương nhau, tuy nhiên chưa thể đánh giá chính xác về mức độ tương tự giữa hai sản phẩm này do tỉ lệ người dùng đánh giá thấp, trong bảng 2.19 ta thấy mức độ tương tự giữa hai sản phẩm này rất cao.

Đối với các cặp sản phẩm i_{24} và i_{34} đều có chung số lượng người dùng đánh giá là 2/7 chiếm 28.5% trong tổng số người dùng với mức đánh giá hoàn toàn khác nhau, như vậy mức độ tương tự giữa hai sản phẩm này thấp, tuy nhiên trong bảng 2.19 ta thấy mức độ tương tự giữa hai sản phẩm này rất cao.

Đối với cặp sản phẩm i_{25} có chung số lượng người dùng đánh giá là 1/7 chiếm 14% trong tổng số người dùng với mức đánh giá hoàn toàn khác nhau, như vậy mức độ tương tự giữa hai sản phẩm này thấp, tuy nhiên trong bảng 2.19 ta thấy mức độ tương tự giữa hai sản phẩm này rất cao.

- **Hệ tương quan Pearson ràng buộc (CPCC)**: Đối với cặp sản phẩm i_{23} có chung số lượng người dùng đánh giá là 1/7 chiếm 14% trong tổng số người dùng với mức đánh giá tương đương nhau, tuy nhiên chưa thể đánh giá chính xác về mức độ tương tự giữa hai sản phẩm này do tỉ lệ người dùng đánh giá thấp, trong bảng 2.19 ta thấy mức độ tương tự giữa hai sản phẩm này rất cao.

Đối với các cặp sản phẩm i_{24} có chung số lượng người dùng đánh giá là 2/7 chiếm 28.5% trong tổng số người dùng và cặp sản phẩm i_{25} có chung số lượng người dùng đánh giá là 1/7 chiếm 14% trong tổng số người dùng với mức đánh giá hoàn toàn khác nhau, như vậy mức độ tương tự giữa các cặp sản phẩm này thấp, tuy nhiên trong bảng 2.19 ta thấy mức độ tương tự giữa hai sản phẩm này rất cao.

2.4. Kết luận

Trong chương này, luận văn đã trình bày về sáu độ đo tương tự sử dụng trong tư vấn lọc cộng tác bao gồm công thức tính toán và ý nghĩa của các ký hiệu sử dụng trong công thức, ví dụ minh họa cách tính các độ đo từ đó đưa ra được những dự đoán phù hợp về xếp hạng cho người dùng hoặc sản phẩm. Vấn đề đặt ra là cần đánh giá các độ đo tương tự sử dụng cùng một thuật toán để xem xét mức độ phù hợp của các độ đo tương tự đó. Ở chương tiếp theo, luận văn sẽ thử nghiệm các độ đo tương tự với thuật toán K-Means trên bộ dữ liệu MovieLens 100K để đưa ra kết quả tư vấn, so sánh và đánh giá các độ đo tương tự áp dụng trong tư vấn lọc cộng tác.

Chương 3 . THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Giới thiệu chung

Để đánh giá các độ đo tương tự có rất nhiều thuật toán phân cụm được sử dụng như: BIRCH, DBSCAN, OPTICS,… Tuy nhiên trong luận văn này tác giả sử dụng thuật toán K-Means để phân cụm đánh giá các độ đo tương tự dựa trên bộ dữ liệu đánh giá của người dùng MovieLens 100K trên website https://grouplens.org/datasets/movielens/.

3.2. Phát biểu bài toán

Lọc cộng tác là lọc luồng dữ liệu mà *Hệ thống đề xuất* (RS) có thể đưa ra cho người dùng mục tiêu theo sở thích của họ. Hồ sơ của người dùng mục tiêu được xây dựng dựa trên sự tương đồng của họ với những người dùng khác. Do đó độ tương đồng là một phần rất quan trọng để đánh giá được mức độ tương tự giữa hai người dùng (hoặc hai sản phẩm). Trong chương 2 luận văn đã trình bày một số độ đo tương tự được sử dụng trong tư vấn lọc cộng tác. Để tìm hiểu rõ hơn trong chương 3 này luận văn sẽ giải quyết bài toán đánh giá các độ đo tương tự như sau:

Input	Ouput
Bộ dữ liệu MovieLens 100k đánh giá của người dùng với các bộ phim trên website: https://grouplens.org/datasets/movielens/ Thuật toán K-Means.	Dữ liệu đã được phân cụm dựa trên các độ đo tương tự: Tương tự Cosine, hệ số tương quan Pearson, hệ số tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, chỉ số Jaccard, khoảng cách Euclide.

Mục đích thử nghiệm của luận văn: Sử dụng thuật toán K-Means để phân cụm dữ liệu sử dụng các độ đo tương tự khác nhau để kiểm tra kết quả phân cụm đầu ra với các độ đo tương tự khác nhau thì dữ liệu phân cụm có sự khác nhau như thế nào. Thuật toán phân cụm K-Means là một trong những thuật toán phân cụm dữ liệu dựa trên học tập không giám sát được sử dụng nhiều trong học máy nói chung và trong

khai phá dữ liệu nói riêng. Ưu điểm của thuật toán là dễ dàng cài đặt và cho kết quả dễ hiểu, linh hoạt trong việc sử dụng các phương pháp đo khoảng cách.

Thuật toán K-Means được phát biểu như sau:

Bước 1: Chọn số k để quyết định số lượng cụm.

Bước 2: Khởi tạo ngẫu nhiên trọng tâm cụm $C=\{c_1,c_2,\ldots, c_k\}$

Bước 3: Lặp

a. Đối với mỗi điểm dữ liệu (x_i) trong tập dữ liệu (D)

- Tính khoảng cách $dis(x_i,C)$ giữa x_i và tất cả các trọng tâm cụm.

- Gán x_i cho cụm gần nhất.

b. Tính toán lại các trọng tâm của cụm làm giá trị trung bình của tất cả các thành viên trong cụm.

Bước 4: Dừng khi các thành viên cụm ổn định.

3.3. Dữ liệu thử nghiệm và phương pháp đánh giá

3.3.1. Mô tả dữ liệu

MovieLens 100K là tập dữ liệu mô tả xếp hạng 5 sao và hoạt động gắn thẻ văn bản miễn phí từ MovieLens, một dịch vụ đề xuất phim. Bộ dữ liệu chứa 100836 xếp hạng và 3683 ứng dụng thẻ trên 9742 phim. Những dữ liệu này được tạo bởi 610 người dùng trong khoảng thời gian từ ngày 29 tháng 3 năm 1996 đến ngày 24 tháng 9 năm 2018. Tập dữ liệu này được tạo vào ngày 26 tháng 9 năm 2018.

Người dùng được chọn ngẫu nhiên để đưa vào. Tất cả những người dùng được chọn đã đánh giá ít nhất 20 phim. Không có thông tin nhân khẩu học được bao gồm. Mỗi người dùng được đại diện bởi một id và không có thông tin nào khác được cung cấp. Dữ liệu được chứa trong các tệp *links.csv, movies.csv, ratings.csv* và *tags.csv*.

ratings.csv: Tất cả các xếp hạng đều có trong tệp *ratings.csv*. Mỗi dòng của tệp này sau hàng tiêu đề đại diện cho một xếp hạng phim của một người dùng và có định dạng sau: *userId, movieId, rating, timestamp*

Các dòng trong tệp này được sắp xếp đầu tiên theo *userId*, sau đó bên trong user được sắp xếp theo *movieId*. Xếp hạng được thực hiện trên thang điểm 5 sao, với số gia tăng nửa sao (0,5 sao - 5,0 sao).

tags.csv: Tất cả các thẻ được chứa trong tệp *tags.csv*. Mỗi dòng của tệp này sau hàng tiêu đề đại diện cho một thẻ được một người dùng áp dụng cho một bộ phim và có định dạng sau: *userId, movieId, tag, timestamp.* Các dòng trong tệp này được sắp xếp đầu tiên theo *userId*, sau đó bên trong user được sắp xếp theo *movieId*.

Thẻ là siêu dữ liệu do người dùng tạo về phim. Mỗi thẻ thường là một từ đơn hoặc cụm từ ngắn. Ý nghĩa, giá trị và mục đích của một thẻ cụ thể được xác định bởi mỗi người dùng.

movies.csv: Thông tin phim được chứa trong tệp *movies.csv*. Mỗi dòng của tệp này sau hàng tiêu đề đại diện cho một bộ phim và có định dạng sau: *movieId, title, genres.* Tên phim được nhập theo cách thủ công hoặc nhập từ trang web https://www.themoviedb.org/ và bao gồm năm phát hành trong ngoặc đơn.

Các thể loại là một danh sách được phân tách bằng dấu sổ đứng và được chọn từ các danh sách sau: Hoạt động, Cuộc phiêu lưu, Hoạt hình, Trẻ em, Phim hài, Tội ác, Phim tài liệu, Kịch, Tưởng tượng, Kinh dị, Âm nhạc, Huyền bí, Lãng mạn,…

links.csv: Các số nhận dạng có thể được sử dụng để liên kết đến các nguồn dữ liệu phim khác được chứa trong tệp *links.csv*. Mỗi dòng của tệp này sau hàng tiêu đề đại diện cho một bộ phim và có định dạng sau: *movieId, imdbId, tmdbId*

- *movieId* là mã định danh cho phim được https://movielens.org sử dụng. Vd: phim Toy Story có link https://movielens.org/movies/1.
- *imdbId* là mã định danh cho phim được sử dụng bởi http://www.imdb.com. Vd: phim Toy Story có link là http://www.imdb.com/title/tt0114709/.
- *tmdbId* là mã định danh cho phim được sử dụng bởi https://www.themoviedb.org. Vd: phim Toy Story có link https://www.themoviedb.org/movie/862.

Trong trong tệp *movies.csv* chọn hai thể loại phim là phim lãng mạn và phim khoa học viễn tưởng để kiểm tra. Trước tiên tính toán xếp hạng trung bình của mỗi người dùng đối với tất cả các phim lãng mạn và tất cả các phim khoa học viễn tưởng. Sau đó loại bỏ những người thích cả khoa học viễn tưởng và lãng mạn để các cụm có xu hướng xác định họ thích thể loại này hơn thể loại khác.

3.3.2. Môi trường và công cụ

Hệ điều hành Windows 11 64bit, RAM 8GB

Phần mềm Visual Studio Code, Jupyter notebook

Ngôn ngữ lập trình: Python

3.4. Cài đặt thuật toán

Cài đặt thuật toán K-Means

- Chọn số lượng cụm

Cách thứ nhất là thử với từng giá trị k=1,2,3,4,5,… để xem kết quả phân cụm thay đổi như thế nào. Một số nghiên cứu cho thấy việc thay đổi k sẽ có hiệu quả nhưng sẽ dừng lại ở một con số nào đó. Như vậy hoàn toàn có thể thử xem dữ liệu tốt với giá trị k nào đó.

Cách thứ 2 sử dụng phương pháp khửu tay (Elbow Method) bằng cách chạy thuật toán K-means nhiều lần, tăng số lượng cụm lên mỗi lần lặp lại. Ghi lại tổn thất cho mỗi lần lặp và sau đó lập biểu đồ đường của các cụm so với tổn thất.

Từ các tính toán theo phương pháp khửu tay dựa vào bộ dữ liệu MovieLens100K ta xác định được k = 2 và x_i là những người dùng.

- Khởi tạo các Centers

```
def _init_centers(self, points, K, **kwargs):
    row, col = points.shape
    retArr = np.empty([K, col])
    for number in range(K):
        randIndex = np.random.randint(row)
        retArr[number] = points[randIndex]

    return retArr
```

- Gán điểm cho Centers gần nhất

```
def _update_assignment(self, centers, points):
    row, col = points.shape
    cluster_idx = np.empty([row])
    distances = self.pairwise_dist(points, centers)
    cluster_idx = np.argmin(distances, axis=1)

    return cluster_idx
```

- Cập nhật Centers mới

```
def _update_centers(self, old_centers, cluster_idx, points):

    K, D = old_centers.shape
    new_centers = np.empty(old_centers.shape)
    for i in range(K):
        new_centers[i] = np.mean(points[cluster_idx == i], axis = 0)
    return new_centers
```

- Hàm mất mát

```
def __call__(self, points, K, max_iters=100, abs_tol=1e-16, rel_tol=1e-16, verbose=False, **kwargs):

    centers = self._init_centers(points, K, **kwargs)
    for it in range(max_iters):
        cluster_idx = self._update_assignment(centers, points)
        centers = self._update_centers(centers, cluster_idx, points)
        loss = self._get_loss(centers, cluster_idx, points)
        K = centers.shape[0]
        if it:
            diff = np.abs(prev_loss - loss)
            if diff < abs_tol and diff / prev_loss < rel_tol:
                break
        prev_loss = loss
        if verbose:
            print('iter %d, loss: %.4f' % (it, loss))
    return cluster_idx, centers, loss
```

- Tìm clusters

```
def find_optimal_num_clusters(data, distance_type, max_K=10):

    y_val = np.empty(max_K)

    for i in range(max_K):
        cluster_idx, centers, labels, y_val[i] = KMeans()(data, i + 1, distance_type)

    plt.plot(np.arange(max_K) + 1, y_val)
    plt.show()
    return y_val
```

- Hiển thị K-Means

```
def kmeans_display(data, label):
    K = np.amax(label) + 1
    X0 = data[label == 0, :]
    X1 = data[label == 1, :]
    X2 = data[label == 2, :]

    plt.plot(X0[:, 0], X0[:, 1], 'b^', markersize = 4, alpha = .8)
    plt.plot(X1[:, 0], X1[:, 1], 'go', markersize = 4, alpha = .8)
    plt.plot(X2[:, 0], X2[:, 1], 'rs', markersize = 4, alpha = .8)

    plt.axis('equal')
    plt.plot()
    plt.show()
```

Cài đặt các độ đo tương tự

- Khoảng cách Euclide

```
def pairwise_dist(self, x, y):  # [5 pts]
    xSumSquare = np.sum(np.square(x),axis=1);
    ySumSquare = np.sum(np.square(y),axis=1);
    mul = np.dot(x, y.T);
    dists = np.sqrt(abs(xSumSquare[:, np.newaxis] + ySumSquare-2*mul))
    return dists
```

- Chỉ số Jaccard

```
def jaccard_similarity(self, x):

    # Calculate all pairwise distances
    jaccard_distances = pdist(x, metric='jaccard')

    # Convert the distances to a square matrix
    jaccard_distances = squareform(jaccard_distances)
    jaccard_similarity = 1-jaccard_distances
    return jaccard_similarity
```

- Tương tự Cosine

```
def cosine_similarity(self, x, y):
    return cosine_similarity(x, y)
```

- Hệ số tương quan Pearson

```
def pearson_cc(self, x, y):
    return np.corrcoef(x, y)
```

- Hệ số tương quan Pearson ràng buộc

```
def constrained_pearson(self, x, y):
    stats.pearsonr(x, y)
```

- Tương quan Pearson dựa trên chức năng Sigmoid

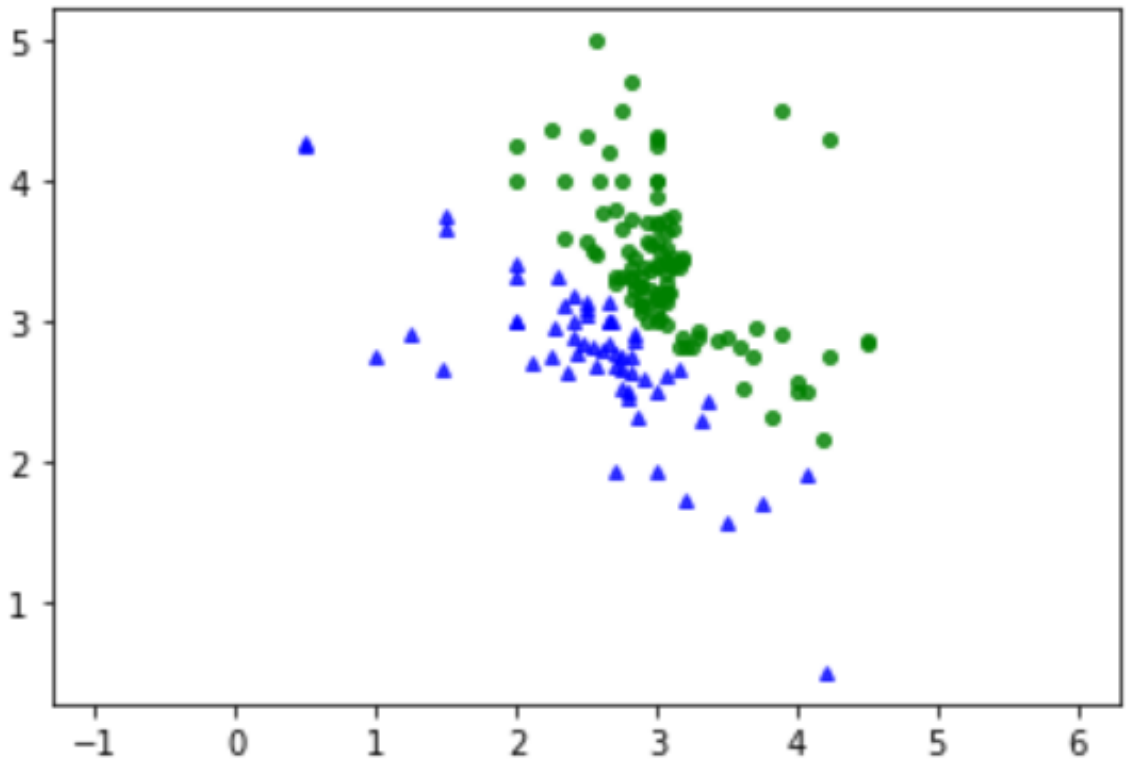
```
def sigmoid(self, x):

    sig = np.where(x < 0, np.exp(x)/(1 + np.exp(x)), 1/(1 + np.exp(-x)))
    return sig
```

3.5. Kết quả thử nghiệm

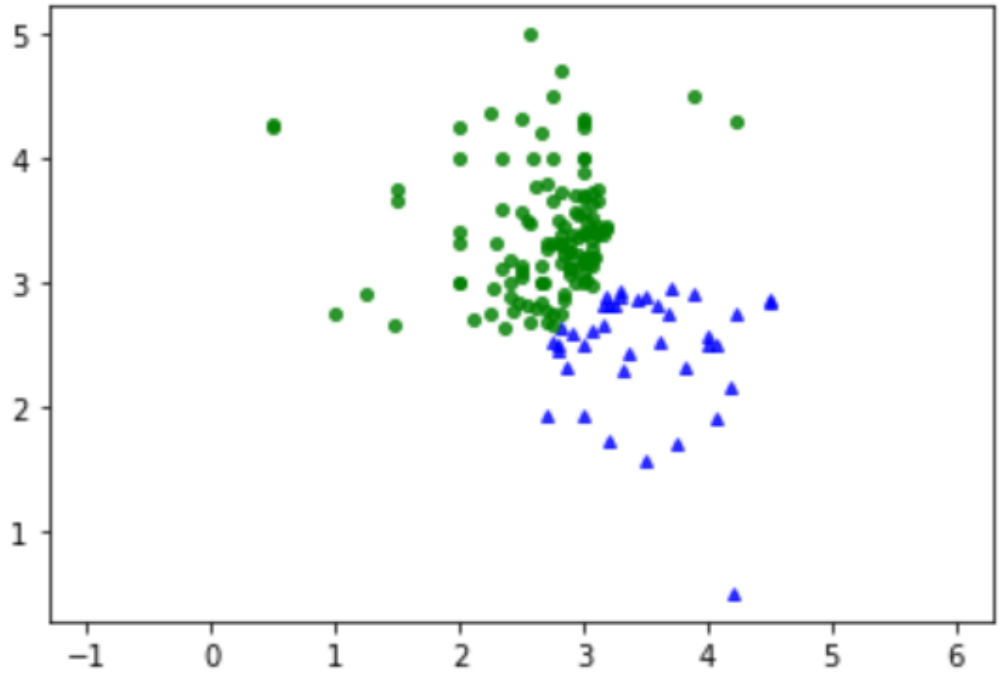
Ta có các kết quả thử nghiệm khi sử dụng thuật toán K-Means với k=2 và dùng các độ đo tương tự khác nhau để phân cụm dữ liệu được thể hiện như sau:

- Khoảng cách Euclide



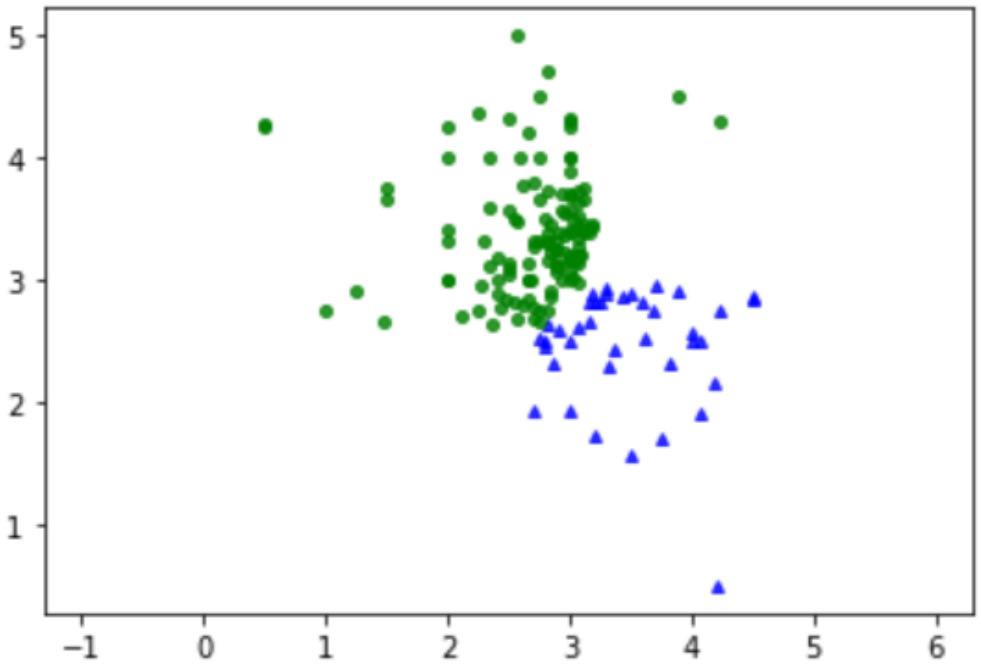
Hình 3.1: Phân cụm sử dụng độ đo tương tự Khoảng cách Euclide

- Tương tự Cosine



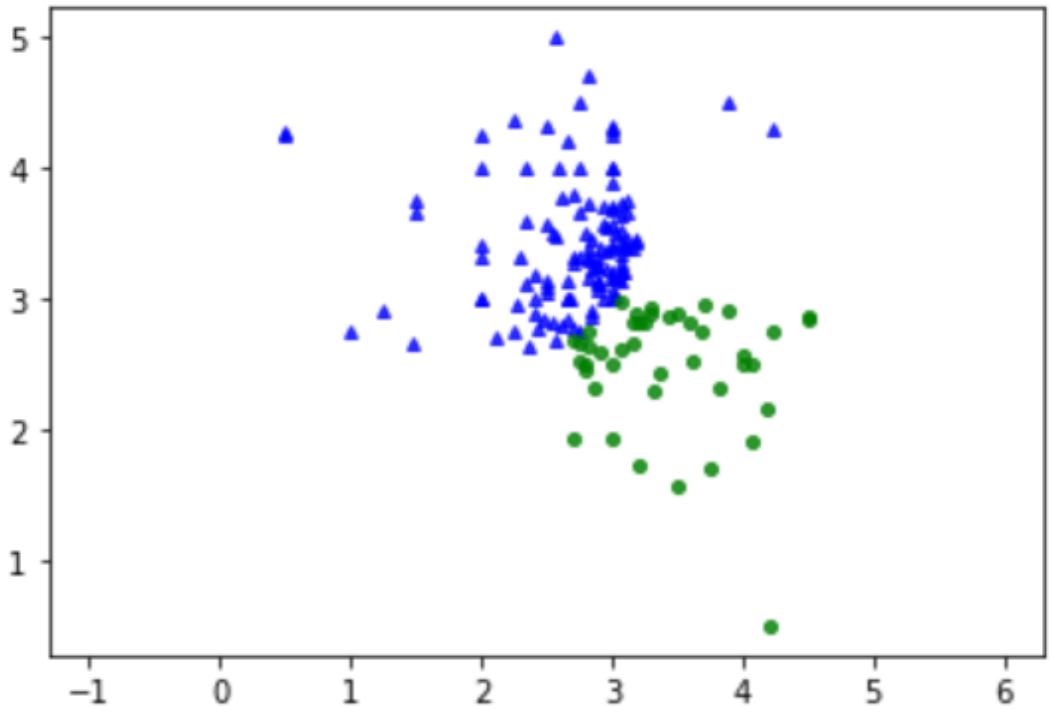
Hình 3.2: Phân cụm sử dụng độ đo tương tự Cosine

- Hệ số tương quan Pearson



Hình 3.3: Phân cụm sử dụng độ đo tương tự Hệ số tương quan Pearson

- Tương quan Pearson dựa trên chức năng Sigmoid



Hình 3.4: Phân cụm sử dụng độ đo Tương quan Pearson dựa trên chức năng Sigmoid

Đánh giá các cụm sử dụng các độ đo khác nhau

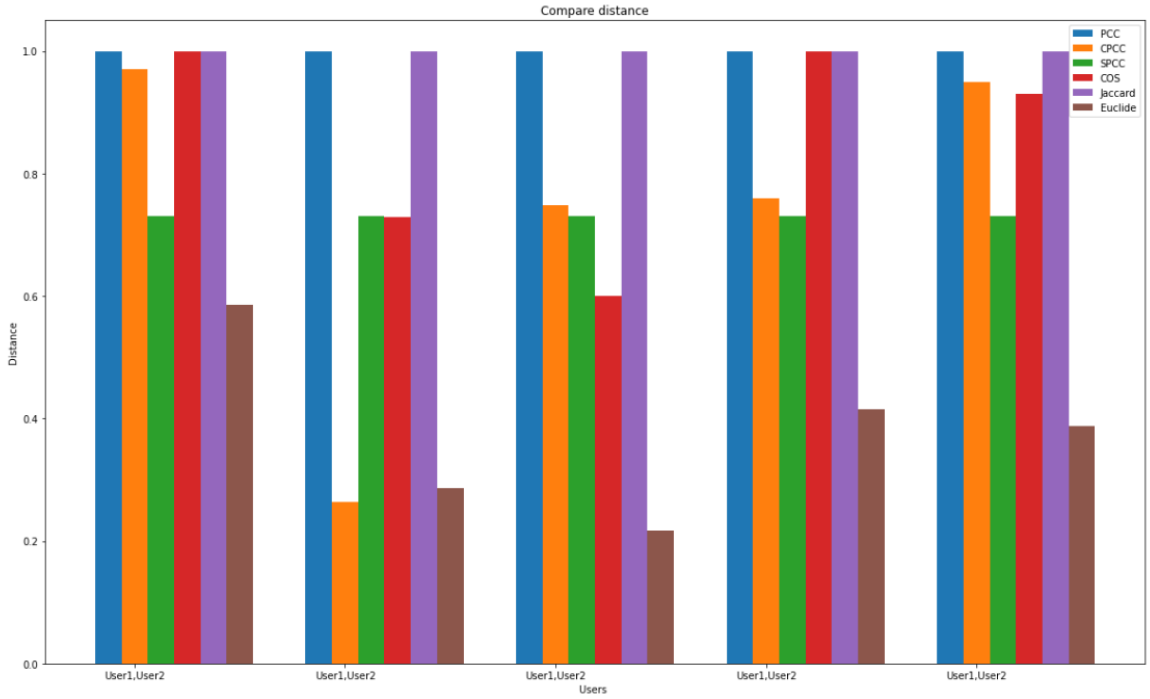
Các cụm dữ liệu sử dụng các độ đo được thể hiện trong các hình: Khoảng cách Euclide (Hình 3.1), Tương tự Cosine (Hình 3.2), Hệ số tương quan Pearson (Hình 3.3), Tương quan Pearson dựa trên chức năng Sigmoid (Hình 3.4). Quan sát các cụm kết quả được đưa ra dựa vào thuật toán K-Means sử dụng các độ đo tương tự trên ta thấy:

Các cụm dữ liệu sử dụng các độ đo Tương tự Cosine, Hệ số tương quan Pearson, Tương quan Pearson dựa trên chức năng Sigmoid cho kết quả tương tự nhau, như vậy kết quả tính khoảng cách giữa các cặp người dùng với các độ đo trên có mức độ chênh lệch thấp không đủ để thay đổi vị trí vào các cụm khác nhau.

Cụm dữ liệu sử dụng Khoảng cách Euclide được phân bố khác so với các cụm sử dụng các độ đo tương tự khác tuy nhiên vẫn có những điểm chung

Đánh giá các các độ đo tương tự

Lấy ngẫu nhiên năm cặp người dùng đưa ra khoảng cách của các cặp người dùng này sử dụng các độ đo tương tự: Tương tự Cosine, hệ số tương quan Pearson, hệ số tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, chỉ số Jaccard, khoảng cách Euclide nhận được kết quả thể hiện trên đồ thị (Hình 3.5)



Hình 3.5: Đồ thị thể hiện độ đo tương tự một số cặp người dùng

Quan sát trên đồ thị (hình 3.5) ta thấy mức độ tương tự giữa cặp người dùng sử dụng các độ đo tương tự khác nhau sẽ cho kết quả khác nhau. Phần lớn các độ đo tương tự cho kết quả giống nhau trong việc tìm kiếm những người dùng.

Trong trường hợp thứ nhất: Mức độ tương tự của hai người dùng với các độ đo PCC, CPCC, COS và J là tương tự nhau; còn lại kết quả với độ đo SPCC và E tương tự nhau

Trong trường hợp thứ 2: Mức độ tương tự của hai người dùng tương tự nhau với các độ đo: PCC và E, SPCC và COS; CPCC và J.

Trong trường hợp thứ 3: Mức độ tương tự của hai người dùng tương tự nhau với các độ đo: PCC và J, CPCC - SPCC và COS; E cho kết quả khác hoàn toàn so với các độ đo khác.

Trong trường hợp thứ 4 và thứ 5: Các độ đo: PCC, CPCC, SPCC, COS và J cho kết quả tương tự nhau, E cho kết quả khác hoàn toàn so với các độ đo khác.

Như vậy rất khó để có thể đánh giá trong việc sử dụng độ đo nào là tốt nhất, có một số ràng buộc trong việc lựa chọn các độ đo: Chẳng hạn như các độ đo Hệ số tương quan Pearson, tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, khoảng cách Euclide và tương tự Cosine chỉ xem xét các sản phẩm chung đã được đánh giá để đo mức độ tương tự, trong khi hệ số Jaccard không chỉ xem xét các sản phẩm chung cùng được đánh giá bởi hai người dùng mà còn xem xét tổng các sản phẩm được đánh giá bởi hai người dùng đó. Ngoài ra việc sử dụng các độ đo cũng phụ thuộc vào mức độ thưa thớt của dữ liệu, trong mỗi trường hợp mức độ thưa khác nhau thì các độ đo sẽ thể hiện ưu điểm và nhược điểm khác nhau.

3.6. Kết luận

Trong chương này, luận văn đã sử dụng thuật toán K-Means với các độ đo tương tự Hệ số tương quan Pearson, tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, khoảng cách Euclide, tương tự Cosine và hệ số Jaccard để phân cụm dữ liệu trên bộ dữ liệu MovieLens 100K và đưa ra các đánh giá so sánh về các cụm dữ liệu cũng như các độ đo tương tự được sử dụng.

KẾT LUẬN VÀ KIẾN NGHỊ

Hệ thống tư vấn lọc cộng tác là một hệ thống đang phát triển trong nhiều lĩnh vực đặc biệt là thương mại điện tử. Hệ thống tư vấn ngày càng hoàn thiện về chất lượng và giảm thời gian xử lý để đáp ứng nhu cầu tư vấn sản phẩm, dịch vụ cho người dùng.

Luận văn đã trình bày về một số độ đo tương tự sử dụng trong lọc cộng tác. Các kết quả đạt được của luận văn như sau:

- Tìm hiểu tổng quan về tư vấn lọc cộng tác, các kỹ thuật lọc cộng tác.
- Tìm hiểu một số độ đo tương tự sử dụng trong tư vấn lọc cộng tác như: Hệ số tương quan Pearson, tương quan Pearson ràng buộc, tương quan Pearson dựa trên chức năng Sigmoid, khoảng cách Euclide, tương tự Cosine và hệ số Jaccard. Đưa ra ví dụ để sử dụng các công thức tính toán độ đo và đánh giá so sánh các độ đo.
- Sử dụng thuật toán K-Means để phân cụm dữ liệu đánh giá các độ đo dựa vào bộ dữ liệu MovieLens 100K.
- So sánh và đánh giá các độ đo tương tự.

Tuy nhiên, luận văn vẫn còn nhiều điểm hạn chế, luận văn chỉ dừng lại ở mức độ nghiên cứu, tìm hiểu. Số lượng các độ đo nghiên cứu chưa đầy đủ. Những hạn chế này đã đưa ra một số hướng mở cho đề tài tiếp tục phát triển như sau:

- Nghiên cứu một số độ đo tương tự khác sử dụng trong lọc cộng tác.
- Sử dụng một số thuật toán khác để đánh giá các độ đo tương tự.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1] Aberger, Christopher R. and caberger, (2014), "Recommender: An Analysis of Collaborative Filtering Techniques".

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, (2001), "Item-Based Collaborative Filtering Recommendation Algorithms", *Proceedings of the 10th international conference on World Wide Web*, 285–295.

[3] FethiFkih, (2021), "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison", Computer and Information Sciences, Volume 33, Issue 8, October 2021.

[4] Hael Al-bashiri, Mansoor Abdullateef Abdulgabber, Awanis Romli, Hasan Kahtan, (2018), "An improved memory-based collaborative filtering method based on the TOPSIS technique".

[5] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu, (2014), "A new user similarity model to improve the accuracy of collaborative filtering", *Knowledge-Based Systems*, Volume 56, 156-166.

[6] Hyung, J. and Ahn, (2008)**, "**A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem", *Elsevier, Information Sciences*, 178: 37–51.

[7] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, (2004), "Evaluating collaborative filtering recommender systems", ACM Trans. Inf. Syst. 22, 1 (January 2004), 5–53.

[8] Jain G., Mahara T., Tripathi K.N, (2020), "A Survey of Similarity Measures for Collaborative Filtering-Based Recommender System", In: Pant M., Sharma T, Verma O., Singla R., Sikander A. (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol 1053. Springer, Singapore. https://doi.org/10.1007/978-981-15-0751-9_32.

[9] Kai Yu, A. Schwaighofer, V. Tresp, Xiaowei Xu and H. -. Kriegel, (2004), "Probabilistic memory-based collaborative filtering," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 56-69.

[10] K. G. Saranya*, G. Sudha Sadasivam and M. Chandralekha, (2016), " Performance Comparison of Different Similarity Measures for Collaborative Filtering Technique", *Indian Journal of Science and Technology*, Volume: 9, Issue: 29, 1-8.

[11] L.A. Hassanieh, C. A. Jaoudeh, J. B. Abdo and J. Demerjian, (2018), "Similarity measures for collaborative filtering recommender systems," 2018 IEEE Middle East and North Africa Communications Conference (MENACOMM), pp. 1-5, doi: 10.1109/MENACOMM.2018.8371003.

[12] N. Mustafa, A. O. Ibrahim, A. Ahmed and A. Abdullah, (2017), "Collaborative filtering: Techniques and applications", 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), pp. 1-6, doi: 10.1109/ICCCCEE.2017.7867668.

[13] Sivaramakrishnan N, Subramaniyaswamy V, Arunkumar S, Renugadevi A, Ashikamai Kk, (2018), "Neighborhood-based approach of collaborative filtering techniques for book recommendation system", *International Journal of Pure and Applied Mathematics*, Volume 119(No. 12), 13241-13250.

[14] Songjie Gong, (2010), "A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering", *Journal of Software 5(7)*, 745-752.

[15] Sondur, S.D., Nayak, S., & Chigadani, A.P, (2016), "Similarity Measures for Recommender Systems: A Comparative Study", *International Journal for Scientific Research and Development, 2*, 76-80.

[16] Xiaoyuan Su and Taghi M. Khoshgoftaar, (2009), "A Survey of Collaborative Filtering Techniques", *Advances in artificial intelligence*, Volume 2009.

[17] Z. Tan and L. He, (2017) "An Efficient Similarity Measure for User-Based Collaborative Filtering Recommender Systems Inspired by the Physical Resonance Principle," in IEEE Access, vol. 5, pp. 27211-27228, doi:10.1109/ACCESS.2017.2778424.