

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Trần Thị Nghĩa

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

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

TÓM TẮT LUẬN VĂN THẠC SỸ

HÀ NỘI – 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. Trần Đình Quế

Phản biện 1:

Phản biện 2:

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

Vào lúc:..... giờ ngày thángnăm 2022

Có thể tìm hiểu luận văn tại:

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

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.

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. 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. 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, Chỉ số Jaccard, Tương tự cosine, Hệ số tương quan Pearson ràng buộc, Tương quan Pearson dựa trên chức năng Sigmoid, Khoảng cách Euclide*

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 LỘC CỘNG TÁC

1.1. Giới thiệu chung

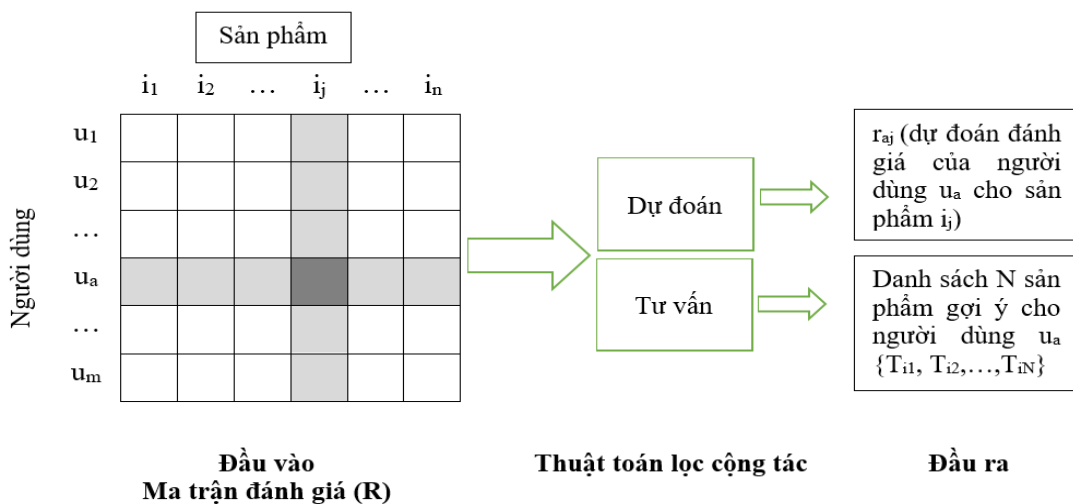
Trong thời đại phát triển của công nghệ thông tin như hiện nay, các trang thương mại điện tử cung cấp lên đến hàng triệu các sản phẩm được bán. Lựa chọn giữa rất nhiều sản phẩm trở thành một công việc đầy thách thức đối với khách hàng. Hệ thống khuyến nghị xuất hiện để giải quyết vấn đề này.

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].

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

Trong một kịch bản CF cổ điển có m là một danh sách người dùng ký hiệu là $U = \{u_1, u_2, \dots, u_m\}$ và n là một danh sách các sản phẩm mà người dùng có thể lựa chọn ký hiệu là $I = \{i_1, i_2, \dots, i_n\}$. 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_i} , 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 $R = \{r_{ij}\}$, $i = 1 \dots m$, $j = 1 \dots n$ 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} . 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

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 thừa 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 thừa thớt và hiệu suất của các dự đoán hoặc khuyến nghị của các hệ thống lọc cộng tác bị thách thức.

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.

1.3.3. Từ đồng nghĩa

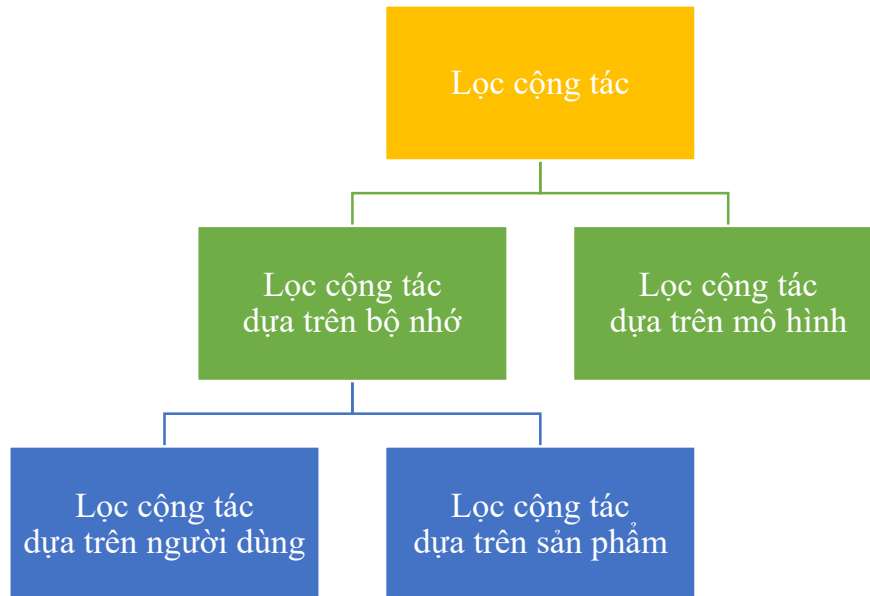
Từ đồng nghĩa đề cập đến xu hướng của một số nội dung giống nhau nhưng có tên nhập khác nhau. Đa số các hệ thống tư vấn không thể phát hiện ra mối liên quan tiềm ẩn này, do đó sẽ xử lý các sản phẩm này một cách khác biệt. Ví dụ: các nội dung có vẻ khác nhau như "*children movie*" và "*children film*" 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.

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

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

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

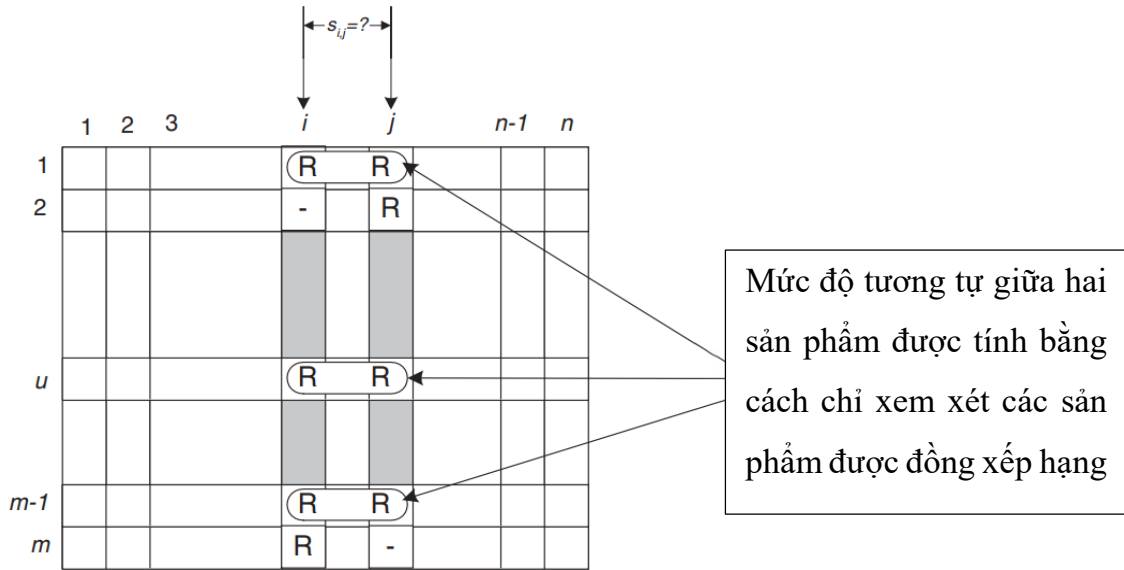
Giải thuật lọc cộng tác dựa trên sản phẩm đề xuất cho người dùng khác với giải thuật lọc cộng tác dựa trên người dùng bởi đối tượng được xét ở đây là các sản phẩm.

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.

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



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

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.

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 Bayes thường được sử dụng cho các nhiệm vụ phân loại.

Đánh giá chưa biết của người dùng u đối với sản phẩm p được tính theo công thức (1.1).

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

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.

1.5. Các tiêu chuẩn đánh giá độ đo

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

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 công thức:

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

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)

RMSE được tính bằng công thức:

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

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

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

- Độ chính xác (P)

$$P = \frac{d}{b+d} \quad (1.6)$$

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)

$$R = \frac{d}{c+d} \quad (1.7)$$

Trong đó:

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

- Độ đo E-measure theo công thức sau:

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

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

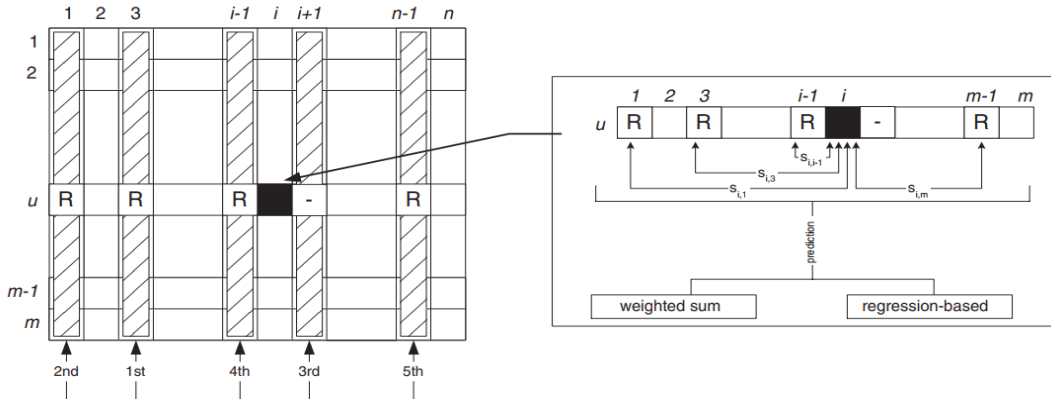
- Độ đo F-measure theo công thức sau:

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

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

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

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)

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

Trong đó

\bar{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

Công thức dự đoán dựa theo sản phẩm được cho như công thức (1.14)

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

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)

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

Trong đó

\bar{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 cộng 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. 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. Chương này sẽ tìm hiểu về một số độ đo tương tự cho tư vấn lọc cộng tác.

2.2. Một số độ đo tương tự

2.2.1. Khoảng cách Euclide (Euclidean distance)

Công thức khoảng cách Euclide biểu thị mối tương quan giữa hai người dùng u và v như sau (20):

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

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 (21):

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

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 Euclidean 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 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)

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).

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

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).

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

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

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

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

Trong đó:

$I_{uv} = \{i \in I | r_{ui} \neq \emptyset \wedge r_{vi} \neq \emptyset\}$ 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 \emptyset của người dùng u

\bar{r}_v : là trung bình cộng các đánh giá khác \emptyset 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:

$U_{ij} = \{u \in U | r_{ui} \neq \emptyset \wedge r_{uj} \neq \emptyset\}$ 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)

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:

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

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)

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 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

2.3.1. Độ tương tự giữa các cặp người dùng

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:

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

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 Cosine

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

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

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

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

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

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

Hệ số tương quan Pearson ràng buộc

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

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

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

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:

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

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 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
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

2.3.2. Độ tương tự giữa các cặp sản phẩm

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:

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

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 Cosine (Cosine similarity)

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

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

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

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

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

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

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 quan Pearson dựa trên chức năng Sigmoid

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

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

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 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

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

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 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 $C=\{c_1, c_2, \dots, c_k\}$

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. Dữ liệu được chứa trong các tệp *links.csv*, *movies.csv*, *ratings.csv* và *tags.csv*.

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

Thực hiện thuật toán K-Means

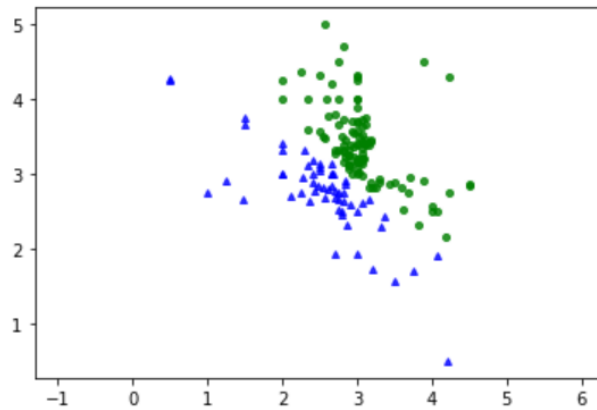
- Khởi tạo các Centers
- Gán điểm cho Centers gần nhất
- Cập nhật Centers mới
- Tìm cluster
- Hiển thị K-Means

Các độ đo tương tự

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

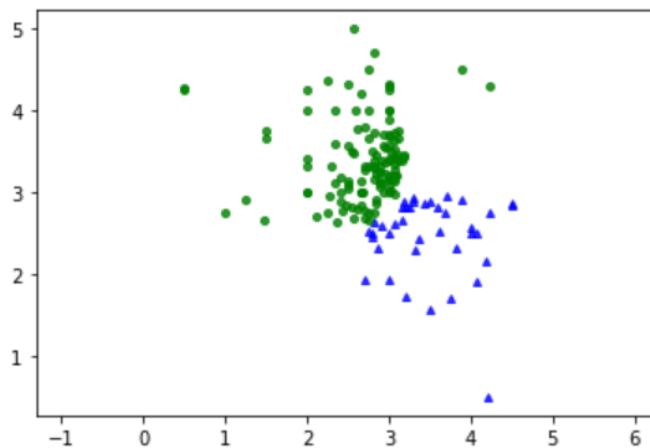
3.5. Kết quả thử nghiệm

- Khoảng cách Euclide



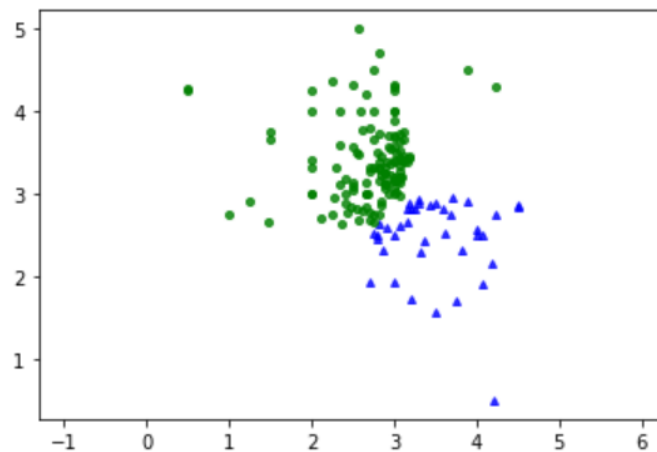
Hình 3.1: Phân cụm sử dụng độ đo tương tự Khoảng cách Euclide

- Tương tự Cosin



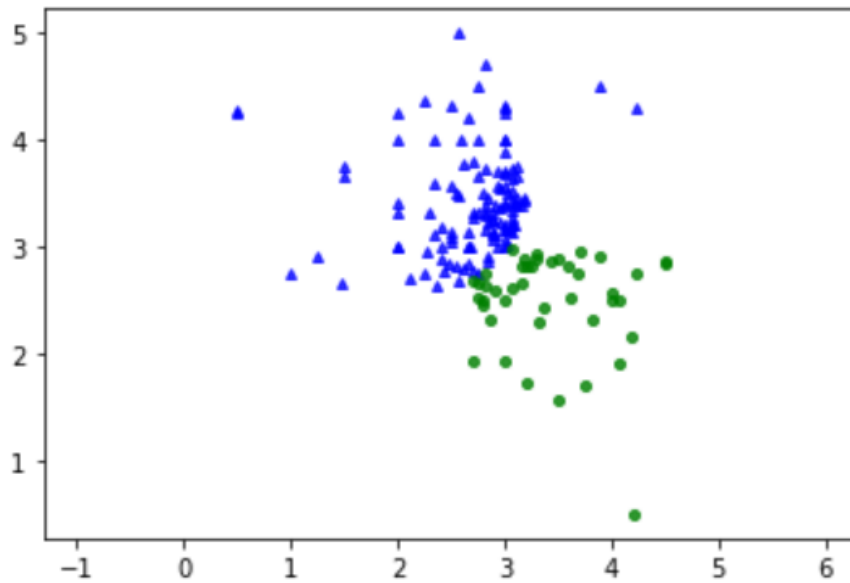
Hình 3.2: Phân cụm sử dụng độ đo tương tự Cosine

- Hệ số tương quan Pearson

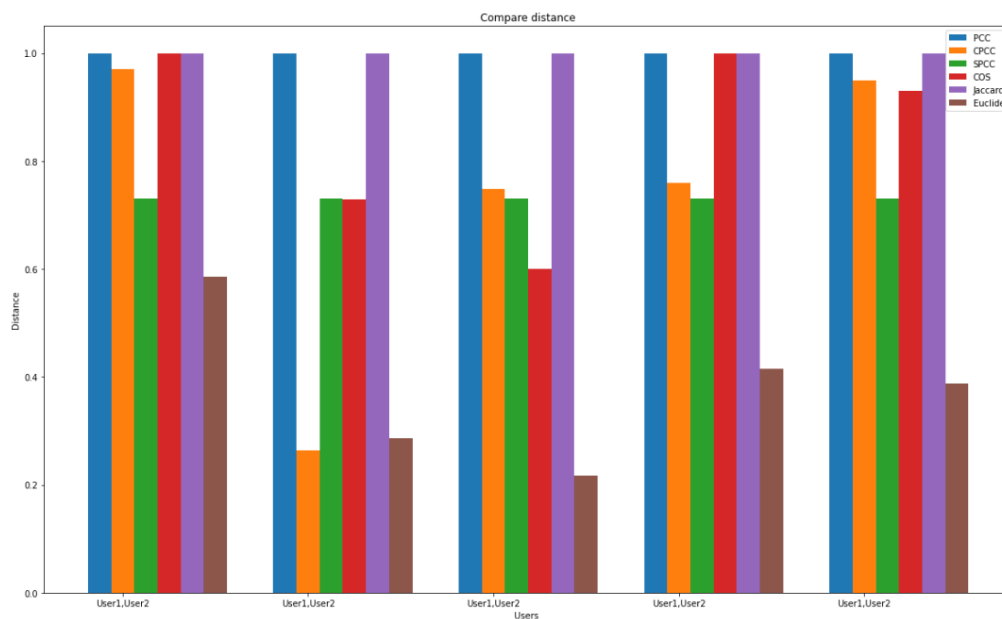


Hình 3.3: Phân cụm sử dụng độ đo tương tự Hệ số tương quan Pearson

- 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



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

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 Euclidean, 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, Subramaniaswamy 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.