BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
****

ĐỖ THỊ LIÊN

PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP XÂY DỰNG HỆ TƯ VẤN

Chuyên ngành: Hệ thống thông tin
Mã số : 9.48.01.04

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2020

Công trình hoàn thành tại:

Học viện Công nghệ Bưu chính Viễn thông

Người hướng dẫn khoa học:

1. GS.TS. Từ Minh Phương

2. TS. Nguyễn Duy Phương

Phản biện 1: ……..

Phản biện 2: ……..

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án tại:

Học viện Công nghệ Bưu chính Viễn thông

Vào hồi: ……giờ, ngày……..tháng…….năm…….

Có thể tìm hiểu luận án tại:

Thư viện Quốc gia Việt Nam

Thư viện Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. Tính cấp thiết của luận án

Với sự gia tăng nhanh chóng của thông tin trên Web thì cần thiết phải có công cụ giúp người dùng lựa chọn các thông tin trực tuyến phù hợp với mình. Để đáp ứng nhu cầu này, các hệ thống tư vấn đã ra đời. Hệ tư vấn (Recommender System) được xem như một hệ thống lọc tích cực, có chức năng hỗ trợ đưa ra quyết định, nhằm mục đích cung cấp cho người sử dụng những gợi ý về thông tin, sản phẩm và dịch vụ phù hợp nhất với yêu cầu và sở thích riêng của từng người tại từng tình huống (ngữ cảnh).

Về cơ bản hệ tư vấn được chia thành hai hướng tiếp cận chính tùy thuộc vào cách khai thác các thông tin đầu vào khác nhau phục vụ cho mục đích tư vấn, đó là: 1) Hệ tư vấn với cách tiếp cận truyền thống; 2) Hệ tư vấn mở rộng cách tiếp cận truyền thống. Trong quá trình nghiên cứu và ứng dụng, mặc dù đã có nhiều nghiên cứu đề xuất được đưa ra để giải quyết bài toán tư vấn theo hai hướng tiếp cận trên, tuy nhiên một số vấn đề mang tính đặc thù đối với thông tin tư vấn như vấn đề dữ liệu thưa, người dùng mới, sản phẩm mới, vấn đề sở thích thay đổi theo thời gian, yêu cầu kết hợp các dạng thông tin khác nhau, làm việc với dữ liệu kích thước lớn được cập nhật thường xuyên… luôn là những vấn đề có tính thời sự và thu hút được sự quan tâm của cộng đồng trong việc nghiên cứu và triển khai vào thực tế.

Đề tài "Phát triển một số phương pháp xây dựng hệ tư vấn" được thực hiện trong khuôn khổ luận án tiến sĩ chuyên ngành hệ thống thông tin nhằm góp phần giải quyết một số vấn đề còn tồn tại trong quá trình xây dựng hệ tư vấn, đó là vấn đề dữ liệu thưa và kết hợp một số dạng thông tin khác nhau vào quá trình tư vấn.

2. Mục tiêu của luận án

Mục tiêu của luận án là nghiên cứu phát triển một số phương pháp xây dựng hệ tư vấn. Đặc biệt, nghiên cứu tập trung vào việc nâng cao độ chính xác của kết quả dự đoán sản phẩm phù hợp với người dùng trong trường hợp dữ liệu thưa, cũng như trong trường hợp có cả dữ liệu sở thích người dùng, thông tin đặc trưng người dùng, thông tin đặc trưng sản phẩm và thông tin ngữ cảnh sử dụng sản phẩm của người dùng. Đồng thời, nghiên cứu cũng tập trung đề xuất một số phương pháp tư vấn đơn giản trong cài đặt để khả thi triển khai thực tế.

3. Các đóng góp của luận án

(1) Đề xuất một phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh.
(2) Đề xuất một phương pháp lọc kết hợp bằng phương pháp đồng huấn luyện.

4. Bố cục của luận án

Chương 1: Tổng quan về hệ tư vấn.
Chương 2: Phát triển phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh.
Chương 3: Phát triển phương pháp lọc kết hợp bằng đồng huấn luyện.

CHƯƠNG 1: TỔNG QUAN VỀ HỆ TƯ VẤN

1.1. Khái niệm hệ tư vấn

Hệ tư vấn, tiếng anh là Recommender System hoặc Recommendation System, là những hệ thống được thiết kế để hướng người dùng đến những đối tượng quan tâm, yêu thích, khi lượng thông tin quá lớn vượt quá khả năng xử lý của người dùng.

Theo Ricci và cộng sự, hệ tư vấn là những công cụ phần mềm, kỹ thuật cung cấp đề xuất các đối tượng có thể hữu ích với người dùng. Những đề xuất liên quan đến quyết định của người dùng như: sản phẩm nào nên mua, bài hát nào nên nghe, hay tin tức nào nên đọc...

1.2. Các lĩnh vực ứng dụng của hệ tư vấn

Hiện tại hệ tư vấn được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, điển hình như thương mại điện tử, giáo dục, giải trí, du lịch, chăm sóc sức khỏe, truyền thông xã hội, ăn uống…

1.3. Phát biểu bài toán tư vấn

Cho tập hợp hữu hạn gồm N người dùng $U = \{u_1, u_1, \dots, u_N\}$ và M sản phẩm $P = \{p_1, p_2, \dots, p_M\}$. Mỗi người dùng $u_i \in U$ (với $i = 1,2, \dots, N$) được biểu diễn thông qua $|T|$ đặc trưng nội dung $T = \{t_1, t_2, \dots, t_{|T|}\}$. Các đặc trưng $t_q \in T$ thông thường là thông tin cá nhân của mỗi người dùng (Demographic Information). Mỗi sản phẩm $p_x \in P$ (với $x = 1,2, \dots M$) 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. Mỗi sản phẩm $p_x \in P$ được biểu diễn thông qua $|C|$ đặc trưng nội dung $C = \{c_1, c_2, \dots, c_{|C|}\}$. Các đặc trưng $c_s \in C$ nhận được từ các phương pháp trích chọn đặc trưng trong lĩnh vực truy vấn thông tin. Mối quan hệ giữa tập người dùng U và tập sản phẩm P được biểu diễn thông qua ma trận đánh giá $R = [r_{ix}]$ với $i = 1, 2, \dots, N; x = 1, 2, \dots, M$ (Hình 1.2).

Sản phẩm

Người dùng	1	2	…	i	…	M
1	5	3	0	1	2	0
2	0	2	0	0	0	4
:	0	0	5	0	0	0
u	3	4	0	2	1	0
:	0	0	0	0	4	0
N	0	0	3	2	0	0
a	3	5	0	?	1	0

Hình 1.2. Ví dụ ma trận đánh giá tổng quát

Gọi $u_a \in U$ là *người dùng hiện thời*, người dùng cần được tư vấn hay *người dùng tích cực*. Khi đó, tồn tại hai dạng bài toán điển hình của hệ tư vấn là:

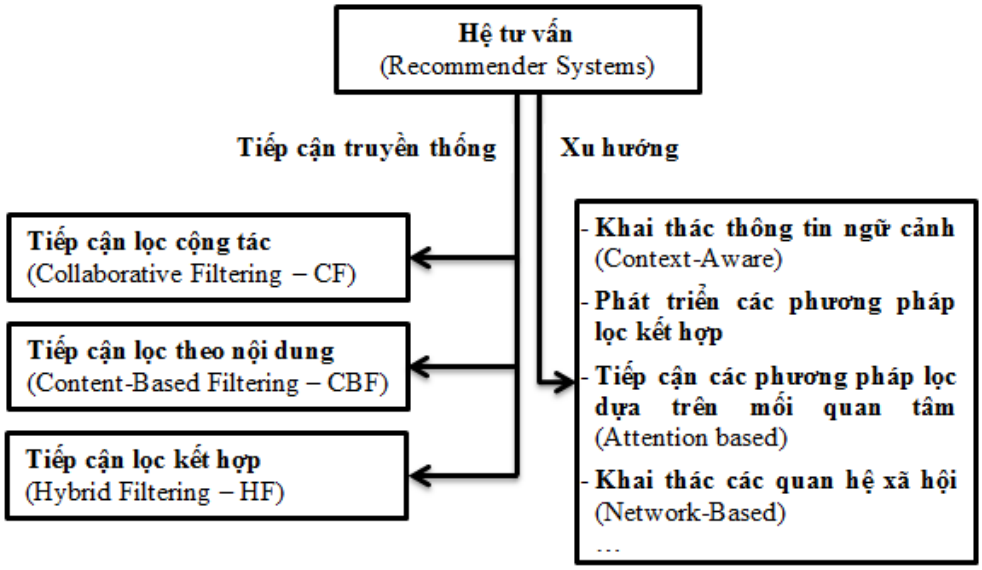
(1) Dự đoán đánh giá của người dùng u_a với các sản phẩm chưa có đánh giá trước đó.

(2) Tư vấn danh sách ngắn các sản phẩm phù hợp với người dùng hiện thời. Cụ thể đối với người dùng u_a, hệ tư vấn sẽ chọn ra K sản phẩm mới $p_x \in (P \backslash P_a)$ phù hợp với người dùng u_a nhất để gợi ý cho họ.

1.4. Qui trình xây dựng hệ tư vấn

Qui trình tổng quát để giải quyết bài toán tư vấn thông thường gồm có 3 giai đoạn chính: 1) Thu thập thông tin; 2) Xây dựng mô hình; 3) Dự đoán đánh giá / Đưa ra tư vấn.

1.5. Các hướng tiếp cận xây dựng hệ tư vấn



Hình 1.4. Các hướng tiếp cận truyền thống và xu hướng hiện nay của hệ tư vấn

1.5.1. Hệ tư vấn sử dụng lọc cộng tác

Lọc cộng tác là phương pháp khai thác những khía cạnh liên quan đến thói quen sử dụng sản phẩm của cộng đồng người dùng có cùng sở thích trong quá khứ để đưa ra dự đoán các sản phẩm mới phù hợp với người dùng hiện thời. Các phương pháp lọc cộng tác nói chung được phân thành hai nhóm chính: 1) *Lọc cộng tác dựa vào bộ nhớ* (Memory-based /Heuristic-based); 2) *Lọc cộng tác dựa vào mô hình* (Model-based). Những vấn đề cần tiếp tục nghiên cứu của lọc cộng tác là vấn đề dữ liệu thưa, vấn đề người dùng mới và sản phẩm mới, vấn đề sở thích thay đổi theo thời gian.

1.5.2. Hệ tư vấn sử dụng lọc nội dung

Lọc theo nội dung là phương pháp gợi ý cho người dùng những sản phẩm mới có nội dung tương tự với các sản phẩm họ đã từng mua hoặc truy nhập trong quá khứ. Các phương pháp tiếp cận cho lọc theo nội dung được chia thành hai nhóm chính: 1) *Lọc nội dung dựa vào bộ nhớ* và 2) *Lọc nội dung dựa vào mô hình*. Những vấn đề cần tiếp tục nghiên cứu của lọc nội dung là vấn đề trích chọn đặc trưng và người dùng mới.

1.5.3. Hệ tư vấn sử dụng lọc kết hợp

Lọc kết hợp là phương pháp kết hợp các kỹ thuật tư vấn khác nhau. Trong đó có bốn xu hướng chính là: 1) *Kết hợp các kết quả dự đoán của lọc cộng tác và lọc nội dung trong lọc kết hợp*; 2) *Kết hợp đặc tính của lọc nội dung vào lọc cộng tác*; 3) *Kết hợp đặc tính của lọc cộng tác vào lọc nội dung*; 4) *Xây dựng mô hình hợp nhất giữa lọc cộng tác và lọc nội dung*. Vấn đề cần tiếp tục nghiên cứu của lọc kết hợp là nâng cao hiệu quả phương pháp biểu diễn và dự đoán cho mô hình kết hợp.

1.5.4. Hệ tư vấn mở rộng cách tiếp cận truyền thống

Các nghiên cứu hiện nay về hệ tư vấn đang tập trung theo hai xu hướng chính: 1) Cải tiến các phương pháp lọc tin truyền thống trong hệ tư vấn; 2) Mở rộng các phương pháp tư vấn truyền thống cho phép tích hợp thêm các nguồn thông tin khác, điển hình là thông tin ngữ cảnh.

1.6. Các phương pháp và độ đo đánh giá hệ tư vấn

1.6.1. Phương pháp đánh giá hệ thống tư vấn

Để đánh giá độ chính xác của hệ thống tư vấn, trước tiên từ ma trận đánh giá R ta tiến hành chia các người dùng (các hàng trong ma trận R) thành hai phần, một phần U_{train} được sử dụng làm dữ liệu huấn luyện, phần còn lại U_{test} được sử dụng để kiểm tra sao cho $U_{train} \cup U_{test} = U$ và $U_{train} \cap U_{test} = \emptyset$. Tập dữ liệu huấn luyện U_{train} được dùng để xây dựng mô hình theo các thuật toán lọc sử dụng trong hệ tư vấn, tập kiểm tra U_{test} được dùng vào quá trình kiểm nghiệm thuật toán tư vấn. Một số cách tiếp cận để chia tập người dùng U thành 2 phần U_{train} *và* U_{test}: Phân chia (Splitting), Lấy mẫu Bootstrap (Bootstrap sampling), Kiểm thử chéo (k-fold cross validation).

1.6.2. Độ đo đánh giá độ chính xác của đánh giá dự đoán

Độ đo điển hình để đánh giá tính chính xác của giá trị dự đoán mà hệ tư vấn đưa ra sẽ căn cứ trên độ sai số giữa giá trị dự đoán và giá trị thực tế. Một số độ đo phổ biến đánh giá sai số phân loại: Độ đo trung bình giá trị tuyệt đối lỗi MAE, độ đo trung bình lỗi lấy căn RMSE.

1.6.3. Độ đo đánh giá độ chính xác của danh sách sản phẩm tư vấn

Một số độ đo phổ biến để đánh giá độ chính xác của danh sách sản phẩm tư vấn: Độ chính xác (Precision), độ nhạy (Recall), E-measure, F-measure; Độ chính xác trung bình tuyệt đối MAP (Mean Average Precision).

1.7. Các nguồn tài nguyên hỗ trợ học tập, nghiên cứu hệ tư vấn

1.8. Kết luận chương 1

Nội dung chương 1 đã trình bày làm rõ khái niệm của hệ tư vấn, phạm vi ứng dụng và phát biểu bài toán hệ tư vấn ở mức tổng quát. Qua đây, luận án phân tích ưu điểm cũng như những mặt còn hạn chế của các phương pháp và nghiên cứu đã có, làm cơ sở để nghiên cứu sinh nghiên cứu phát triển một số phương pháp tư vấn. Các đề xuất của luận án được trình bày trong chương 2 và 3.

CHƯƠNG 2: PHÁT TRIỂN PHƯƠNG PHÁP LỌC CỘNG TÁC DỰA TRÊN MÔ HÌNH ĐỒ THỊ CHO HỆ TƯ VẤN THEO NGỮ CẢNH

2.1. Đặt vấn đề

Một trong số khó khăn chính mà các phương pháp lọc cộng tác gặp phải là vấn đề dữ liệu thưa. Để giải quyết vấn đề dữ liệu thưa cho lọc cộng tác, 2 hướng tiếp cận điển hình được đưa ra: 1) Giảm số chiều của ma trận đánh giá; 2) Khai thác các mối liên hệ gián tiếp trên ma trận đánh giá. Trong chương này luận án trình bày đề xuất một phương pháp mới tính toán mức độ tương tự giữa các cặp người dùng hoặc sản phẩm dựa trên mô hình đồ thị, theo hướng tiếp cận thứ 2. Trên cơ sở độ đo tương tự dựa trên mô hình đồ thị đề xuất cho hệ tư vấn cộng tác với cách tiếp cận truyền thống đưa ra trong Mục 2.2, luận án phát triển hệ tư vấn cộng tác theo ngữ cảnh trong mục 2.3. Mô hình đồ thị cho phép khai thác các mối quan hệ trực tiếp và bắc cầu giữa các đỉnh giúp giải quyết vấn đề dữ liệu thưa, đồng thời khắc phục nhược điểm của các phương pháp cùng hướng trước đó.

2.2. Độ đo tương tự cho lọc cộng tác dựa trên mô hình đồ thị

2.2.1. Biểu diễn đồ thị cho lọc cộng tác

Hệ lọc cộng tác với ma trận đánh giá gồm N người dùng $U = \{u_1, u_1, \dots, u_N\}$ và M sản phẩm $P = \{p_1, p_2, \dots, p_M\}$ hình thành nên một đồ thị hai phía, một phía là tập người dùng, phía

còn lại là tập sản phẩm, ký hiệu là đồ thị $G =< V, E >$. Tập đỉnh V của đồ thị được chia thành hai tập: tập đỉnh người dùng và tập đỉnh sản phẩm ($V = U \cup P$). Tập cạnh E của đồ thị được xác định theo công thức (2.2). Mỗi cạnh $e_{ij} \in E$ kết nối từ đỉnh người dùng u_i tới đỉnh sản phẩm p_j nếu tồn tại đánh giá biết trước của u_i với p_j, có dạng $e = (u_i, p_j)$. Không tồn tại các cạnh của nối giữa hai đỉnh người dùng hoặc cạnh nối giữa hai đỉnh sản phẩm. Trọng số của mỗi cạnh e_{ij} là w_{ij} được xác định theo (2.3).

$$E = \{e = (u_i, p_j): u_i \in U, p_j \in P \,|r_{ij} \neq 0\} \quad (2.2)$$

$$w_{ij} = \begin{cases} r_{ij} & If\ (u_i, p_j) \in E \\ 0 & Otherwise \end{cases} \quad (2.3)$$

2.2.2. Độ đo tương tự cho lọc cộng tác dựa trên biểu diễn đồ thị

2.2.2.1. Độ đo tương tự giữa các cặp người dùng cho lọc cộng tác dựa trên biểu diễn đồ thị

Mức độ tương tự giữa người dùng $u_i \in U$ và người dùng $u_j \in U$ được ước lượng bằng tổng các trọng số của tất cả các đường đi độ dài L đi từ đỉnh u_i đến đỉnh u_j trên đồ thị, với trọng số của mỗi đường đi được tính bằng tích trọng số các cạnh tương ứng. Việc làm này được xác định thông qua ma trận trọng số tổng quát biểu diễn đồ thị G dưới đây.

$$Z = \begin{pmatrix} UZ(N \times N) & W(N \times M) \\ W^T(M \times N) & PZ(M \times M) \end{pmatrix} \quad (2.4)$$

Khi đó, mức độ tương tự giữa các cặp người dùng được tính toán dựa vào ma trận trọng số Z theo công thức sau:

$$\mathrm{UZ^L} = \begin{cases} \mathrm{W.W^T}, & \mathrm{L} = 2 \\ \mathrm{W.W^T.UZ^{L-2}}, & \mathrm{L} = 4,6,8,\ldots \end{cases} \quad (2.5)$$

Định lý 2.1 dưới đây sẽ cho ta một cách xác định L trong trường hợp đồ thị biểu diễn của lọc cộng tác $G =< V, E >$ liên thông.

Định lý 2.1. *Nếu đồ thị biểu diễn cho các hệ lọc cộng tác $G =< V, E >$ liên thông thì luôn luôn tồn tại số tự nhiên chẵn L để $uz_{ij}^L \neq 0$ với mọi $u_i, u_j \in U$. Trong đó, uz_{ij}^L xác định theo (2.5).*

2.2.2.2. Độ đo tương tự giữa các cặp người dùng cho lọc cộng tác dựa trên biểu diễn đồ thị

Mức độ tương tự giữa các cặp sản phẩm được tính toán theo công thức (2.6) sau:

$$PZ^L = \begin{cases} W^T.W, & L = 2 \\ W^T.W.PZ^{L-2}, & L = 4,6,8,\ldots \end{cases} \quad (2.6)$$

Định lý 2.2. *Nếu đồ thị biểu diễn cho các hệ lọc cộng tác $G =< V, E >$ liên thông thì luôn luôn tồn tại số tự nhiên chẵn L để $pz_{xy}^L \neq 0$ với mọi $p_x, p_y \in P$. Trong đó, pz_{xy}^L xác định theo (2.6).*

2.3. Lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh

2.3.1. Ngữ cảnh

Định nghĩa ngữ cảnh: "Thông tin ngữ cảnh là những thông tin có thể mô tả được hoàn cảnh của một thực thể. Thực thể ở đây có thể là người, là vật hoặc là đối tượng có liên quan tới sự tương tác giữa người dùng và ứng dụng, bao gồm cả bản thân người dùng và ứng dụng đó".

2.3.2. Bài toán tư vấn theo ngữ cảnh

Bài toán tư vấn truyền thống được biểu diễn dựa trên ma trận đánh giá hai chiều sau:

$$R_o: U \times P \rightarrow R \quad (2.7)$$

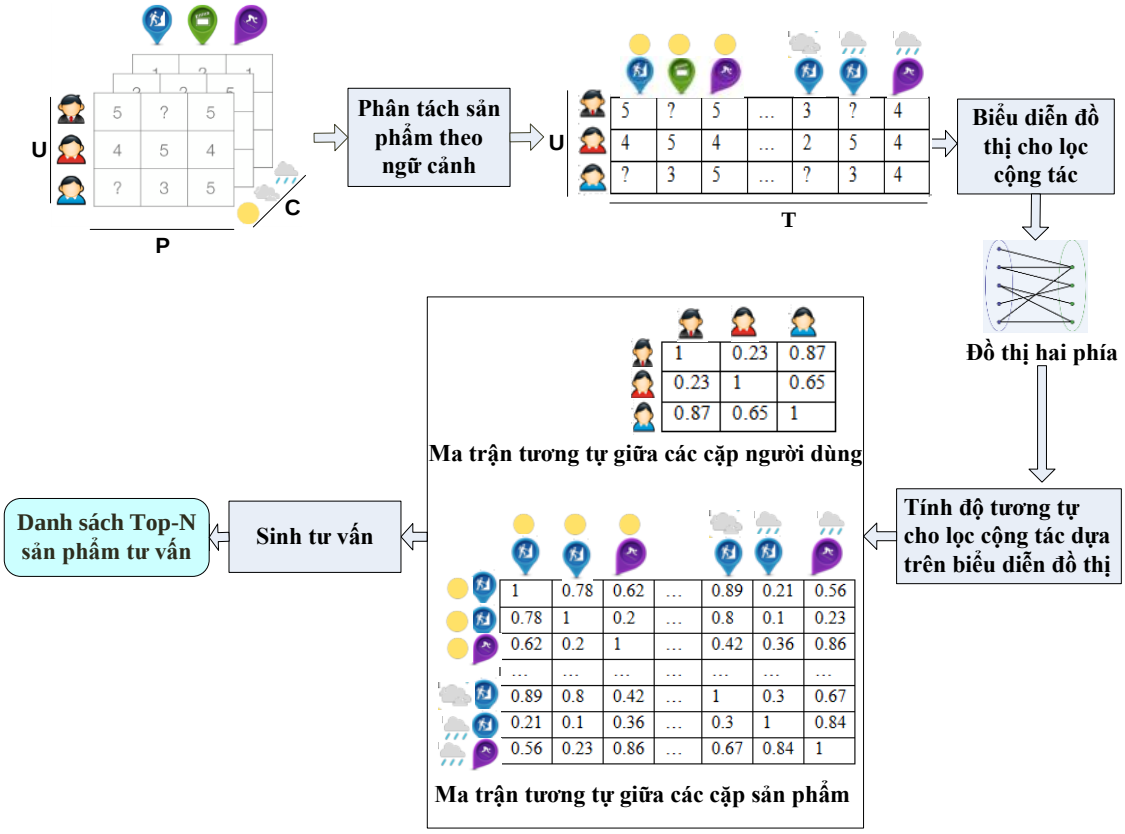
Bài toán tư vấn theo ngữ cảnh sẽ dựa trên ma trận đánh giá đa chiều (Multi-dimensional matrix) như sau:

$$R_1: U \times P \times C \rightarrow R \quad (2.8)$$

Tổng quát hóa, giả sử ta có tập hữu hạn $U = \{u_1, u_1, \dots, u_N\}$ là tập gồm N người dùng, $P = \{p_1, p_2, \dots, p_M\}$ là tập gồm M sản phẩm và K chiều ngữ cảnh $C_1, C_2, \dots, C_K$, mỗi chiều ngữ cảnh có tương ứng $N_{c_1}, N_{c_2}, \dots, N_{c_K}$ điều kiện ngữ cảnh. Mối quan hệ giữa tập người dùng U, tập sản phẩm P và tập ngữ cảnh C được biểu diễn thông qua công thức (2.8). Nhiệm vụ của hệ tư vấn theo ngữ cảnh là dự đoán đánh giá và đưa ra tư vấn các sản phẩm mới cho người dùng trong tình huống ngữ cảnh cụ thể.

2.3.3. Các hướng tiếp cận giải quyết bài toán tư vấn theo ngữ cảnh

Các cách tiếp cận để sử dụng thông tin về ngữ cảnh trong quá trình tư vấn có thể được phân thành 3 hướng tiếp cận: 1) Lọc trước theo ngữ cảnh; 2) Lọc sau theo ngữ cảnh và 3) Mô hình hóa ngữ cảnh. Luận án đề xuất một phương pháp tư vấn cộng tác theo ngữ cảnh mới thuộc hướng tiếp cận lọc trước ngữ cảnh theo hình 2.4 sau.



Hình 2.4. Bộ khung triển khai phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh

2.3.4.1. Phân tách sản phẩm theo ngữ cảnh

Phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến cho phép tích hợp đầy đủ thông tin ngữ cảnh trong việc chuyển hóa sản phẩm ban đầu thành sản phẩm giả lập. Các bước thực hiện cụ thể như sau:

- **Bước 1.** Tạo ra 1 chiều ngữ cảnh mới C đại diện cho K chiều ngữ cảnh $C_1, C_2, \dots, C_K$ bằng cách lấy tích Đề-các của tất cả các chiều ngữ cảnh.
- **Bước 2.** Tạo ra tập sản phẩm giả lập T bằng cách lấy tích Đề-các của tập sản phẩm P và chiều ngữ cảnh C.
- **Bước 3.** Chuyển đổi ma trận đánh giá đa chiều về ma trận đánh giá hai chiều bằng việc loại bỏ đi tập ngữ cảnh, thay tập sản phẩm ban đầu P bằng tập sản phẩm giả lập T.

Quá trình phân tách sản phẩm theo ngữ cảnh sẽ biến đổi ma trận đánh giá đa chiều R_1 (biểu diễn đánh giá của người dùng với sản phẩm trong các tình huống ngữ cảnh khác nhau) về ma trận đánh giá hai chiều R_o (biểu diễn đánh giá của người dùng với sản phẩm giả lập). Để hạn chế những vấn đề dữ liệu thưa của lọc cộng tác áp dụng cho ma trận đánh giá hai chiều R_o, luận án sử dụng phương pháp tính toán toán mức độ tương tự giữa các cặp người dùng hoặc sản phẩm dựa trên mô hình đồ thị đề xuất trong Mục 2.2.

2.3.4.2. Biểu diễn đồ thị cho lọc cộng tác

Áp dụng phương pháp biểu diễn đồ thị cho lọc cộng tác đề xuất trong Mục 2.2.1 cho ma trận đánh giá hai chiều R_o thu được sau bước 2.3.4.1.

2.3.4.3. Tính độ tương tự cho lọc cộng tác dựa trên biểu diễn đồ thị

Việc tính toán mức độ tương tự cho lọc cộng tác dựa vào biểu diễn đồ thị nêu trên đươc chia thành 2 cách tiếp cận theo đề xuất trong 2.2.2.

2.3.4.4. Sinh tư vấn

Áp dụng phương pháp kNN để sinh danh sách các sản phẩm tư vấn phù hợp với người dùng hiện thời với độ đo tương tự trình bày trong Mục 2.3.4.3.

Trên cơ sở bộ khung triển khai phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh, luận án đề xuất hai thuật toán mới cho hệ tư vấn cộng tác theo ngữ cảnh là: 1) Thuật toán lọc cộng tác theo ngữ cảnh dựa vào mức độ tương tự giữa các cặp người dùng trên mô hình đồ thị (IS-UserBased-Graph); 2) Thuật toán lọc cộng tác theo ngữ cảnh dựa vào mức độ tương tự giữa các cặp sản phẩm trên mô hình đồ thị (IS-ItemBased-Graph).

Đầu vào:

- *Ma trận đánh giá đa chiều R_1 (chứa thông tin ngữ cảnh).*
- *$u_a \in U$ là người dùng hiện thời cần được tư vấn.*
- *$c \in (C_1 \times C_2 \times \dots \times C_K)$ là ngữ cảnh ứng với người dùng hiện thời.*
- *K_1 là số lượng người dùng trong tập láng giềng với u_a.*
- *K_2 là số lượng sản phẩm cần tư vấn cho u_a.*

Đầu ra:

- *Danh sách K_2 sản phẩm tư vấn tới người dùng u_a trong tình huống ngữ cảnh c.*

Các bước thực hiện:

Bước 1. *Chuyển đổi ma trận đánh giá dạng đa chiều R_1 về dạng hai chiều R_o*

Theo phương pháp phân tách sản phẩm theo ngữ cảnh (Mục 2.3.4.1).

Bước 2. *Tính mức độ tương tự giữa các cặp người dùng dựa trên mô hình đồ thị*

Biểu diễn đồ thị cho hệ tư vấn (Mục 2.3.4.2).

$L \Leftarrow 2$; //Thiết lập độ dài đường đi ban đầu giữa các cặp người dùng

Repeat

$$UZ^L = \begin{cases} W.W^T, & L = 2 \\ W.W^T.UZ^{L-2}, & L = 4,6,8,\ldots \end{cases}$$

$L \Leftarrow L + 2$; // Tăng độ dài đường đi.

Until ($uz_{ij}^L \neq 0$ *với mọi* $u_j \in (U \backslash u_i)$);

- **Bước 3**. *Sinh tư vấn cho người dùng hiện thời* u_a *trong ngữ cảnh* c.
  - Với mỗi người dùng hiện thời u_a, chọn K_1 người dùng có mức độ tương tự cao nhất với u_a làm tập láng giềng. Kí hiệu U_a là tập láng giềng của u_a gồm K_1 người dùng.
  - Dự đoán đánh giá chưa biết r_{aj} của người dùng u_a với sản phẩm $t_j \in T$

$$r_{aj} = \frac{\sum_{r_{ij} \in R_j} r_{ij}}{|R_j|}, R_j = \{r_{ij} | r_{ij} \neq 0, u_i \in U_a\}$$

  - Chuyển đổi ma trận dự đoán đánh giá hai chiều chứa sản phẩm giả lập (trong tập T) về ma trận dự đoán đánh giá đa chiều chứa sản phẩm thực (thuộc tập P) và tình huống ngữ cảnh đi kèm (thuộc tập C).
  - Chọn K_2 sản phẩm thực trong P có đánh giá dự đoán cao nhất để tư vấn cho người dùng u_a trong tình huống ngữ cảnh c.

Thuật toán 2.1. Thuật toán IS-UserBased-Graph

Đầu vào:

- *Ma trận đánh giá đa chiều* R_1 *(chứa thông tin ngữ cảnh).*
- $u_a \in U$ *là người dùng hiện thời cần được tư vấn.*
- $c \in (C_1 \times C_2 \times \ldots \times C_K)$ *là ngữ cảnh ứng với* u_a.
- K_1 *là số lượng sản phẩm trong tập láng giềng với sản phẩm được* u_a *đánh giá.*
- K_2 *là số lượng sản phẩm cần tư vấn cho* u_a.

Đầu ra:

- *Danh sách* K_2 *sản phẩm tư vấn tới người dùng* u_a *trong tình huống ngữ cảnh* c.

Các bước thực hiện:

Bước 1. *Chuyển đổi ma trận đánh giá dạng đa chiều* R_1 *về dạng hai chiều* R_o

Theo phương pháp phân tách sản phẩm theo ngữ cảnh (Mục 2.3.4.1).

Bước 2. *Tính mức độ tương tự giữa các cặp sản phẩm dựa trên mô hình đồ thị*

Biểu diễn đồ thị cho hệ tư vấn (Mục 2.3.4.2).

$L \Leftarrow 2$; //Thiết lập độ dài đường đi ban đầu giữa các cặp sản phẩm

Repeat

$$TZ^L = \begin{cases} W^T.W, & L = 2 \\ W^T.W.TZ^{L-2}, & L = 4,6,8,\ldots \end{cases}$$

$L \Leftarrow L + 2$; // Tăng độ dài đường đi.

Until ($tz_{kj}^L \neq 0$ *với mọi* $t_k \in (T \backslash t_j)$);

- **Bước 3**. *Sinh tư vấn cho người dùng hiện thời* u_a *trong ngữ cảnh* c.
  - Thực hiện lặp: với mỗi sản phẩm giả lập $t_j \in T$ chưa được đánh giá bởi người dùng u_a
    - Chọn K_1 sản phẩm có mức độ tương tự cao nhất với t_j làm tập láng giềng. Kí hiệu T_j là tập láng giềng của t_j gồm K_1 sản phẩm.
    - Dự đoán đánh giá chưa biết r_{aj} của người dùng u_a với $t_j \in T_j$

$$r_{aj} = \frac{\sum_{r_{ak} \in R_a} r_{ak}}{|R_a|},\ R_a = \{r_{ak} | r_{ak} \neq 0, t_k \in T_j\}$$

- Chuyển đổi ma trận dự đoán đánh giá hai chiều chứa sản phẩm giả lập (trong tập T) về ma trận dự đoán đánh giá đa chiều chứa sản phẩm thực (thuộc tập P) và tình huống ngữ cảnh đi kèm (thuộc tập C).
- Chọn K_2 sản phẩm thực trong P có đánh giá dự đoán cao nhất để tư vấn cho người dùng u_a trong tình huống ngữ cảnh c.

Thuật toán 2.2. Thuật toán IS-ItemBased-Graph

2.4. Thực nghiệm và kết quả

2.4.1. Dữ liệu thực nghiệm

Sử dụng ba bộ dữ liệu DepaulMovie, MovieLens 100K, InCarMusic. Trong đó: *DepaulMovie* chứa 5043 đánh giá từ 97 người dùng cho 79 phim trong các tình huống ngữ cảnh khác nhau, bộ dữ liệu này có 3 chiều ngữ cảnh; *MovieLens 100K* chứa 100000 đánh giá từ 973 người dùng, 1682 phim trong các tình huống ngữ cảnh khác nhau, bộ dữ liệu này có 2 chiều ngữ cảnh; *InCarMusic* chứa 3938 đánh giá từ 1042 người dùng, 139 album trong các tình huống ngữ cảnh khác nhau, bộ dữ liệu này có 8 chiều ngữ cảnh.

2.4.2. Cài đặt thực nghiệm

- **Độ đo:** Precision@N, MAP@N (N=10).
- **Phương pháp thực nghiệm:** Phương pháp kiểm thử chéo (k-fold cross-validation) với k=10. Việc thực nghiệm được thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.
- **Các phương pháp tư vấn được sử dụng để so sánh:** *BiasedMF, UserSplitting-BiasedMF, ItemSplitting-BiasedMF, UISplitting-BasedMF, SLIM, CSLIM, ItemSplitting-SLIM, UserBased-Graph, ItemBased-Graph, ItemSplitting-UserBased-Graph, ItemSplitting-ItemBased-Graph, IS-UserBased-Graph, IS-ItemBased-Graph, IS-Graph.*

2.4.3. Kết quả thực nghiệm

Bảng 2.7. Giá trị Precision@10, MAP@10 trên tập DepaulMovie

Phương pháp	Precision@10	MAP@10
BiasedMF	0.082	0.141
UserSplitting-BiasedMF	0.089	**0.162**
ItemSplitting-BiasedMF	0.086	0.147
UISplitting-BiasedMF	0.084	0.144
SLIM	0.084	0.145
CSLIM	0.085	0.121
ItemSplitting-SLIM	0.092	0.158
UserBased-Graph	0.087	0.149
ItemBased-Graph	0.085	0.150
ItemSplitting-UserBased-Graph	0.122	0.134
ItemSplitting -ItemBased-Graph	0.124	0.151
IS-UserBased-Graph	0.121	0.159
IS-ItemBased-Graph	**0.125**	0.158
IS-Graph	0.117	0.148

Bảng 0.1. Giá trị Precision@10, MAP@10 trên tập MovieLens 100K

Phương pháp	Precision@10	MAP@10
BiasedMF	0.027	0.0064
UserSplitting-BiasedMF	0.030	0.0076
ItemSplitting-BiasedMF	0.029	0.0065
UISplitting-BiasedMF	0.028	0.0066
SLIM	0.022	0.0060
CSLIM	0.004	0.0005
ItemSplitting-SLIM	0.023	0.0061
UserBased-Graph	0.028	0.0065
ItemBased-Graph	0.034	0.0068
ItemSplitting-UserBased-Graph	0.057	0.0085
ItemSplitting -ItemBased-Graph	0.069	0.0097
IS-UserBased-Graph	0.085	0.0104
IS-ItemBased-Graph	**0.103**	**0.0108**
IS-Graph	0081	0.0089

Bảng 0.2. Giá trị Precision@10, MAP@10 trên tập InCarMusic

Phương pháp	Precision@10	MAP@10
BiasedMF	0.032	0.121
UserSplitting-BiasedMF	0.033	0.125
ItemSplitting-BiasedMF	0.034	0.127
UISplitting-BiasedMF	0.033	0.117
SLIM	0.023	0.064
CSLIM	0.018	0.038
ItemSplitting-SLIM	0.023	0.065
UserBased-Graph	0.033	0.123
ItemBased-Graph	0.035	0.130
ItemSplitting-UserBased-Graph	0.035	0.063
ItemSplitting -ItemBased-Graph	0.036	0.111
IS-UserBased-Graph	0.034	**0.147**
IS-ItemBased-Graph	**0.037**	0.142
IS-Graph	0.014	0.115

Một số nhận xét được đưa ra căn cứ vào phân tích kết quả thực nghiệm như sau:

1) Các phương pháp lọc cộng tác cho hệ tư vấn không sử dụng ngữ cảnh: Việc khai thác mối quan hệ bắc cầu giữa các đỉnh dựa vào mô hình đồ thị giúp cải thiện đáng kể chất lượng dự đoán của *UserBased-Graph, ItemBased-Graph* so với các phương pháp cơ sở trong các hệ tư vấn không sử dụng ngữ cảnh.
2) Các phương pháp phân tách theo ngữ cảnh (UserSplitting / ItemSplitting / UISplitting) kết hợp với phương pháp phân rã ma trận MF cho chất lượng tư vấn tốt hơn phương pháp *BiasedMF* thuần túy cho lọc cộng tác. Điều này hoàn toàn phù hợp với những nghiên cứu trước đây [113].
3) Các phương pháp phân tách theo ngữ cảnh kết hợp với phương pháp *BiasedMF* cho chất lượng tư vấn tốt hơn phương pháp *CSLIM* trên cả ba tập dữ liệu. Phương pháp *CSLIM* cho độ chính xác thấp hơn phương pháp *ItemSplitting-SLIM*, thậm chí thấp hơn *SLIM* ở 2 trong 3 tập dữ liệu. Điều đó cho thấy sự kết hợp của các phương pháp phân tách theo ngữ cảnh với các phương

pháp tư vấn truyền thống cho lại hiệu quả tư vấn khá tốt so với các phương pháp tư vấn theo ngữ cảnh khác, đây cũng là hướng tiếp cận để đưa ra đề xuất phương pháp tư vấn theo ngữ cảnh mới của tác giả trong luận án.

4) Các phương pháp dựa trên mô hình đồ thị sử dụng 1 chiều ngữ cảnh *ItemSplitting-UserBased-Graph, ItemSplitting-ItemBased-Graph* cho lại Precision@10 tốt hơn, nhưng MAP@10 lại cho kết quả thấp hơn các phương pháp dựa trên mô hình đồ thị không sử dụng ngữ cảnh *UserBased-Graph* / *ItemBased-Graph* và phương pháp tư vấn theo ngữ cảnh cơ sở cùng hướng sử dụng kết hợp *ItemSplitting.* Như vậy có thể khẳng định việc dùng 1 chiều ngữ cảnh trong phương pháp phân tách sản phẩm theo ngữ cảnh kết hợp với phương pháp dựa trên đồ thị chưa hẳn là giải pháp tối ưu.

5) Việc sử dụng đồng thời nhiều chiều ngữ cảnh giúp bổ sung thông tin hữu ích cho quá trình tư vấn hơn việc sử dụng 1 chiều ngữ cảnh xét cả ở tiêu chí Precision@10 và MAP@10. Kết quả kiểm nghiệm cũng chỉ ra rằng phương pháp đề xuất *IS-UserBased-Graph, IS-ItemBased-Graph* cho lại độ chính xác $Precision@10$ tốt hơn các phương pháp cơ sở. Đặc biệt, phương pháp IS-ItemBased-Graph cho $Precision@10$ cao nhất đối với cả ba tập dữ liệu và $MAP@10$ cao nhất trên tập dữ liệu MovieLens. Phương pháp *IS-UserBased-Graph* cho $MAP@10$ cao nhất trên tập dữ liệu InCarMusic. Quan sát riêng trên tập dữ liệu DepaulMovie, tác giả nhận thấy phương pháp *UserSplitting-BiasedMF* cho MAP@10 cao nhất các phương pháp khác, điều này có thể được lý giải là do DepaulMovie là tập dữ liệu ít thưa thớt nhất trong ba tập dữ liệu. Các kết quả này đưa ra bằng chứng cho thấy phương pháp đề xuất bởi luận án ít nhạy cảm với dữ liệu thưa thớt so với các phương pháp tư vấn theo ngữ cảnh cơ sở, dù thực tế phương pháp đề xuất tích hợp đầy đủ các thông tin ngữ cảnh.

   Trong hai phương pháp đề xuất bởi luận án, *IS-ItemBased-Graph* cho độ chính xác $Precision@10$ cao hơn *IS-UserBased-Graph*, điều này được lý giải là bởi vì tại bước 1 của thuật toán, các sản phẩm được phân tách thành các sản phẩm giả lập nên thông tin về sản phẩm được khai thác chi tiết và đầy đủ hơn cho quá trình huấn luyện và sinh tư vấn sau đó.

6) Phương pháp đề xuất bởi luận án IS-UserBased-Graph, IS-ItemBased-Graph cho lại độ chính xác cao hơn IS-Graph, điều đó có thể khẳng định việc kết hợp khai thác mối quan hệ bắc cầu giữa các cặp người dùng hoặc các cặp sản phẩm và giải thuật kNN cho lại hiệu quả tư vấn tốt hơn việc khai thác mối quan hệ bắc cầu giữa đỉnh người dùng và sản phẩm trên đồ thị trước đây.

2.5. Kết luận chương 2

Chương này đã trình bày một độ đo tương tự giữa các cặp người dùng hoặc các cặp sản phẩm mới để giải quyết bài toán lọc cộng tác cho hệ tư vấn truyền thống và trọng tâm vào mở rộng cho hệ tư vấn theo ngữ cảnh. Phương pháp lọc cộng tác dựa trên mô hình đồ thị đề xuất cho hệ tư vấn theo ngữ cảnh cho phép tích hợp đầy đủ thông tin ngữ cảnh vào quá trình dự đoán sản phẩm phù hợp cho người dùng và hạn chế ảnh hưởng vấn đề thưa dữ liệu đánh giá. Kết quả kiểm nghiệm trên cả ba tập dữ liệu thực cho thấy phương pháp đề xuất cho lại kết quả dự đoán tốt hơn các phương pháp tư vấn theo ngữ cảnh cơ sở, đặc biệt trong trường hợp dữ liệu thưa.

CHƯƠNG 3: PHÁT TRIỂN PHƯƠNG PHÁP LỌC KẾT HỢP BẰNG ĐỒNG HUẤN LUYỆN

3.1. Đặt vấn đề

Lọc kết hợp là phương pháp kết hợp các phương pháp tư vấn khác nhau cho phép ta tận dụng được lợi thế mỗi phương pháp trong việc nâng cao kết quả dự đoán. Trong chương này, luận án tiếp cận hướng kết hợp đặc tính của lọc nội dung vào lọc cộng tác dựa vào bộ nhớ để phát triển phương pháp lọc kết hợp mới cho hệ tư vấn. Mục 3.2 trình bày đề xuất một phương pháp mới giải quyết vấn đề dữ liệu thưa cho lọc cộng tác bằng đồng huấn luyện. Trên cơ sở lọc cộng tác bằng phương pháp đồng huấn luyện, luận án đề xuất phương pháp lọc kết hợp mới bằng đồng huấn luyện ở Mục 3.3 nhằm giải quyết vấn đề dữ liệu và tích hợp hiệu quả các đặc trưng nội dung vào lọc cộng tác.

3.2. Lọc cộng tác bằng phương pháp đồng huấn luyện

Bài toán lọc cộng tác nhằm dự đoán các đánh giá chưa biết từ tập các đánh giá đã biết có thể phát biểu như bài toán phân lớp cơ sở của học máy.

3.2.1. Phát biểu bài toán lọc cộng tác bằng phân lớp

Nhiệm vụ của lọc cộng tác là điền vào hay dự đoán các giá trị thích hợp cho các giá trị chưa có đánh giá trong ma trận đánh giá. Tiếp cận lọc cộng tác bằng phân lớp ta cần cá nhân hóa mô hình học theo người dùng hoặc theo sản phẩm nhằm gán nhãn cho những giá trị đánh giá chưa biết trong ma trận đánh giá. Các nhãn này thuộc cùng dải giá trị với các giá trị đánh giá đã biết.

3.2.2. Phân lớp bằng phương pháp đồng huấn luyện

3.2.2.1. Giải quyết bài toán phân lớp theo hướng tiếp cận học bán giám sát

Xét mức độ phù hợp của các hướng tiếp cận học máy cho hệ tư vấn, với thông tin đầu vào là ma trận đánh giá, tác giả nhận định rằng: Với ma trận đánh giá ban đầu chỉ có một số rất ít đánh giá biết trước, để có thể khai thác đầy đủ dữ liệu gán nhãn và chưa gán nhãn từ ma trận đánh giá đầu vào cho hệ tư vấn nhằm hạn chế ảnh hưởng của vấn đề dữ liệu thưa, tác giả tập trung nghiên cứu vào hướng tiếp cận học bán giám sát cho bài toán phân lớp, trong trường hợp này là bài toán lọc cộng tác.

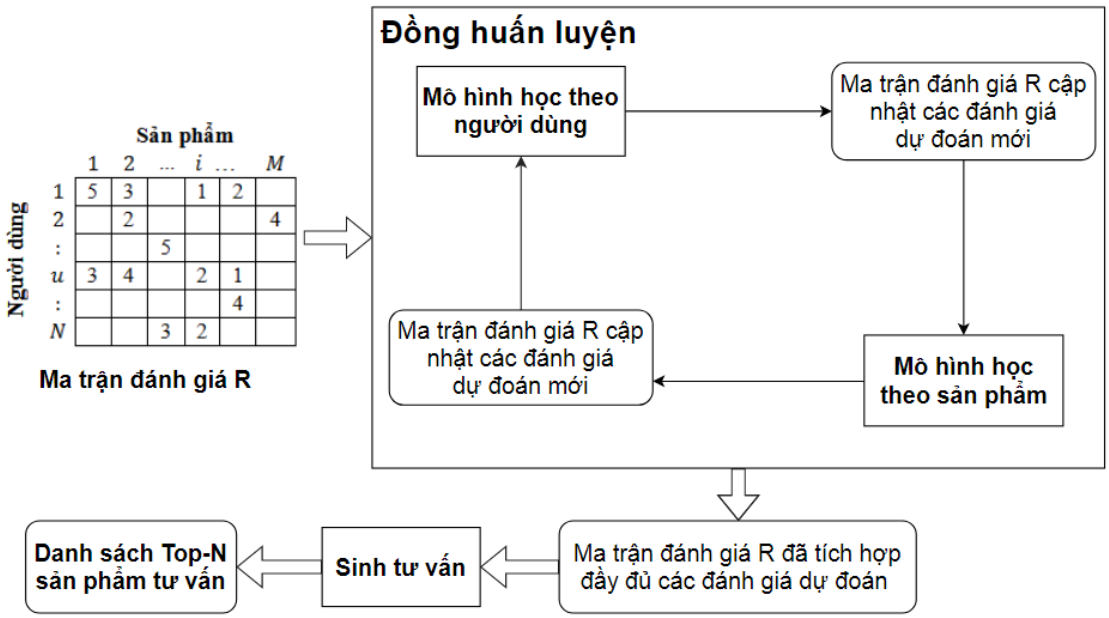
3.2.2.2. Phát biểu bài toán phân lớp bằng học bán giám sát

Cho tập hữu hạn D^L gồm các mẫu dữ liệu đã được gán nhãn, $D^L = \{x_i, y_i\}_i^L$ và tập hữu hạn D^U gồm các mẫu dữ liệu chưa được gán nhãn, $D^U = \{x_j\}_{j=L+1}^{L+U}$. Nhiệm vụ của bài toán phân lớp dữ liệu là cần xây dựng một mô hình phân lớp để khi có một mẫu dữ liệu mới vào thì mô hình phân lớp sẽ cho biết mẫu dữ liệu đó thuộc lớp nào. Với hướng tiếp cận học bán giám sát cho bài toán phân lớp thì cả hai tập dữ liệu đã được gán nhãn và chưa được gán nhãn ở trên đều tham gia vào việc huấn luyện và dự đoán lớp. Trong phạm vi luận án, tác giả đề xuất một cách tiếp cận dựa vào phương pháp đồng huấn luyện cho bài toán phân lớp của lọc cộng tác.

3.2.2.3. Bán giám sát bằng phương pháp đồng huấn huyện

Phương pháp đồng huấn luyện được đánh giá là phù hợp cho các bộ dữ liệu chứa các mẫu dữ liệu được quan sát dưới hai góc nhìn độc lập nhau, khi đó phương pháp này cho phép 2 bộ phân lớp học riêng biệt trên mỗi góc nhìn dữ liệu và kết hợp các dự đoán để giảm lỗi phân lớp. Quá trình này được lặp lại đến khi thỏa mãn điều kiện các mẫu dữ liệu đều được gán nhãn hoặc số vòng lặp đạt đến ngưỡng xác định trước.

3.2.3. Mô hình đồng huấn luyện cho lọc cộng tác



Hình 3.1. Bộ khung triển khai lọc cộng tác bằng phương pháp đồng huấn luyện

3.2.3.1. Mô hình học theo người dùng

Việc xác định mức độ tương tự giữa các cặp người dùng $u_i \in U$ không dùng để xác định tập láng giềng K_i tác động trực tiếp lên tư vấn như trong, mà chỉ để dùng vào việc xác định các nhãn phân loại chắc chắn r_{iy} cho người dùng u_i. Để thực hiện điều này, tác giả đưa ra khái niệm *tập sinh cho người dùng* $u_i \in U$ theo định nghĩa 3.1 dưới đây.

Định nghĩa 3.1. *Tập sinh cho người dùng* $u_i \in U$ *được ký hiệu là* S_i *là tập tất cả những người dùng* $u_j \in U$ *có đánh giá giao nhau với* u_i *tối thiểu* γ *sản phẩm. Trong đó,* γ *là hằng số nguyên dương.*

$$S_i = \{u_j \in U : |P_i \cap P_j| \geq \gamma\} \quad (3.1)$$

Mức độ tương tự của mỗi mỗi người dùng $u_i \in U$ và người dùng $u_j \in U$ chỉ được tính toán trên tập sinh $S_i \in U$.

$$u_{ij} = \begin{cases} 0 & \text{If } u_j \notin S_i \\ \dfrac{\sum_{p_x \in P_i \cap P_j}(r_{ix} - \bar{r_i})(r_{jx} - \bar{r_j})}{\sqrt{\sum_{p_x \in P_i \cap P_j}(r_{ix} - \bar{r_i})^2 \sum_{p_x \in P_i \cap P_j}(r_{jx} - \bar{r_j})^2}} & \text{, Otherwise} \end{cases} \quad (3.2)$$

Tập láng giềng của người dùng $u_i \in U$ được xác định theo định nghĩa 3.2 dưới đây.

Định nghĩa 3.2. *Tập láng giềng của người dùng* $u_i \in U$*, ký hiệu* K_i*, là tập những người dùng* u_j *thuộc tập sinh* S_i *có mức độ tương tự* u_{ij} *được xác định theo công thức (3.2) vượt quá ngưỡng* β*. Trong đó,* $\beta \in [0,1]$.

$$K_i = \{u_j \in S_i \,|u_{ij} > \beta\} \quad (3.3)$$

Dựa trên tập láng giềng K_i của người dùng $u_i \in U$, các mẫu dữ liệu chưa có đánh giá được gán nhãn giá trị dự đoán (nhãn phân loại chắc chắn) theo công thức (3.4).

$$r_{ix} = \overline{r_i} + \frac{\sum_{u_j \in K_i}(r_{jx} - \overline{r_j})u_{ij}}{\sum_{u_j \in K_i}|u_{ij}|} \quad (3.4)$$

3.2.3.2. Mô hình học theo sản phẩm

Tương tự như đối với người dùng, việc xác định mức độ mức độ tương tự giữa các cặp sản phẩm dựa trên khái niệm tập sinh cho sản phẩm $p_x \in P$ theo định nghĩa 3.3 dưới đây.

Định nghĩa 3.3. *Tập sinh cho sản phẩm $p_x \in P$ được ký hiệu là C_x là tập tất cả sản phẩm $p_y \in P$ có đánh giá giao nhau với p_x tối thiểu γ người dùng. Trong đó, γ là hằng số nguyên dương.*

$$C_x = \{p_y \in P : |U_x \cap U_y| \geq \gamma\} \quad (3.5)$$

Mức độ tương tự của mỗi sản phẩm $p_x \in P$ và sản phẩm $p_y \in P$ chỉ được tính toán trên tập sinh $C_x \in P$.

$$p_{xy} = \begin{cases} 0 & \text{If } p_y \notin C_x \\ \dfrac{\sum_{u_i \in U_x \cap U_y}(r_{ix} - \overline{r_x})(r_{iy} - \overline{r_y})}{\sqrt{\sum_{u_i \in U_x \cap U_y}(r_{ix} - \overline{r_x})^2 \sum_{u_i \in U_x \cap U_y}(r_{ix} - \overline{r_y})^2}} & \text{, Otherwise} \end{cases} \quad (3.6)$$

Tập láng giềng của sản phẩm $p_x \in P$ được xác định theo định nghĩa 3.4 dưới đây.

Định nghĩa 3.4. *Tập láng giềng của sản phẩm $p_x \in P$ được ký hiệu là K_x là tập những sản phẩm p_y thuộc tập sinh C_x có mức độ tương tự p_{xy} được xác định theo công thức (3.6) vượt quá ngưỡng β. Trong đó, $\beta \in [0,1]$.*

$$K_x = \{p_y \in C_x \mid p_{xy} > \beta\} \quad (3.7)$$

Dựa trên tập láng giềng K_x của sản phẩm $p_x \in P$, nhãn phân loại chắc chắn cho người dùng $u_i \in U$ được dự đoán theo công thức (3.8).

$$r_{ix} = \frac{\sum_{p_y \in K_x} p_{xy} r_{iy}}{\sum_{p_y \in K_x}|p_{xy}|} \quad (3.8)$$

3.2.3.2. Lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng

Đầu vào: *Khởi tạo ma trận đánh giá* $R^{(0)} = \left\{r_{ix}^{(0)}\right\} = \{r_{ix}\}$.

Đầu ra : *Ma trận dự đoán* $R^{(t)} = \left\{r_{ix}^{(t)}\right\}$.

Các bước tiến hành:

1. Khởi tạo số bước lặp ban đầu: $t \leftarrow 0$;
2. Bước lặp:

Repeat

2.1. Tăng bước lặp: $t \leftarrow t + 1$;

2.2. Huấn luyện theo người dùng:

a) *Tìm* $S_i^{(t)}, u_{ij}^{(t)}$ *theo công thức* (3.1), (3.2)

b) *Tìm* $K_i^{(t)}$ *theo công thức* (3.3).

c) *Dự đoán* $r_{ix}^{(t)}$ *theo công thức* (3.4).

2.3. Huấn luyện theo sản phẩm:

a) *Tìm* $C_x^{(t)}, p_{xy}^{(t)}$ *theo công thức* (3.5), (3.6).

b) *Tìm* $K_x^{(t)}$ *theo công thức* (3.7).

c) *Dự đoán* $r_{ix}^{(t)}$ *theo công thức* (3.8).

Until $(r_{ix}^{(t)} = r_{ix}^{(t-1)})$

Thuật toán 3.2. Thuật toán CoTrainning-UserItem.

Tính hội tụ và điều kiện cần và đủ để thuật toán CoTraining-UserItem có thể điền đầy đủ các giá trị dự đoán theo mệnh đề 3.1 và định lý 3.1 dưới đây.

Mệnh đề 3.1. *Thuật toán CoTraining-UserItem sẽ hội tụ tại vòng lặp thứ t khi không có nhãn phân loại nào được bổ sung vào ma trận dự đoán, khi đó* $r_{ix}^{(t)} = r_{ix}^{(t-1)}$ *với* $i = 1,2,\dots,N; x = 1,2,\dots,M$.

Định lý 3.1. *Điều kiện cần và đủ để dự đoán quan điểm của người dùng* $u_i \in U$ *cho tất cả các sản phẩm mới* $p_x \in P$ *một giá trị đánh giá* $r_{ix} \neq 0$ *theo phương pháp CoTraining-UserItem là* $\bigcup_{u_j \in K_i} P_j = P$. *Trong đó,* K_i *được xác định theo công thức (3.3).*

3.2.3.3. Lọc cộng tác bằng phương pháp đồng huấn luyện theo sản phẩm

Đầu vào:*Khởi tạo ma trận đánh giá* $R^{(0)} = \left\{r_{ix}^{(0)}\right\} = \{r_{ix}\}$.

Đầu ra :*Ma trận dự đoán* $R^{(t)} = \left\{r_{ix}^{(t)}\right\}$.

Các bước tiến hành:

1. Khởi tạo số bước lặp ban đầu: $t \leftarrow 0$;

2. Bước lặp:

Repeat

2.1. Tăng bước lặp: $t \leftarrow t + 1$;

2.2. Huấn luyện theo sản phẩm:

a) *Tìm* $C_x^{(t)}, p_{xy}^{(t)}$ *theo công thức* (3.5), (3.6).

b) *Tìm* $K_x^{(t)}$ *theo công thức* (3.7).

c) *Dự đoán* $r_{ix}^{(t)}$ *theo công thức* (3.8).

2.3. Huấn luyện theo người dùng:

a) *Tìm* $S_i^{(t)}, u_{ij}^{(t)}$ *theo công thức* (3.1), (3.2).

b) *Tìm* $K_i^{(t)}$ *theo công thức* (3.3).

c) *Dự đoán* $r_{ix}^{(t)}$ *theo công thức* (3.4).

Until $(r_{ix}^{(t)} = r_{ix}^{(t-1)})$

Thuật toán 3.3. Thuật toán CoTraining-ItemUser

Tính hội tụ và điều kiện cần và đủ để thuật toán CoTraining-ItemUser có thể điền đầy đủ các giá trị dự đoán theo mệnh đề 3.2 và định lý 3.2 dưới đây.

Mệnh đề 3.2. *Thuật toán CoTraining-ItemUser sẽ hội tụ tại vòng lặp thứ t khi không có nhãn phân loại nào được bổ sung vào ma trận dự đoán, khi đó* $r_{ix}^{(t)} = r_{ix}^{(t-1)}$ *với* $i = 1,2,\dots,N; x = 1,2,\dots,M$.

Định lý 3.2. *Điều kiện cần và đủ mỗi người dùng* $u_i \in U$ *đều được dự đoán các sản phẩm mới* $p_x \in P$ *một giá trị đánh giá* $r_{ix} \neq 0$ *là* $\bigcup_{p_y \in K_x} U_y = U$. *Trong đó,* K_x *được xác định theo công thức (3.7).*

3.2.3.2. Sinh tư vấn

Từ ma trận $R^{(t)}$ thu được sau quá trình đồng huấn luyện, tiến hành sắp xếp các sản phẩm chưa được đánh giá ban đầu bởi người dùng hiện thời u_a theo thứ tự giảm dần của $r_{ix}^{(t)}$. Sau đó, chọn K sản phẩm đầu tiên trong số đó tư vấn cho người dùng u_a.

3.3. Lọc kết hợp bằng phương pháp đồng huấn luyện

3.3.1. Hợp nhất biểu diễn giá trị các đặc trưng nội dung vào ma trận đánh giá

3.3.1.1. Hợp nhất hồ sơ người dùng của lọc nội dung vào ma trận đánh giá

Gọi $P_i \subseteq P$ được xác đinh theo (3.12) là tập sản phẩm $p_x \in P$ đã được đánh giá bởi người dùng $u_i \in U$.

$$P_i = \{p_x \in P \mid r_{ix} \neq 0 \ \ (u_i \in U)\} \tag{3.12}$$

Gọi $\mathrm{Item}(i,s)$ là tập các sản phẩm $p_x \in P_i$ chứa đựng đặc trưng $c_s \in C$ được xác định theo công thức (3.13).

$$Item(i,s) = \{p_x \in P_i \mid c_{xs} \neq 0 \ \ (u_i \in U,\ c_s \in C)\} \tag{3.13}$$

Dựa trên P_i và $\mathrm{Item}(i,s)$ các phương pháp tư vấn theo nội dung ước lượng được trọng số w_{is} phản ánh mức độ quan trọng của đặc trưng nội dung c_s đối với người dùng u_i. Trong đề xuất này tác giả đưa ra một phép trích chọn đặc trưng có cùng mức độ đánh giá tự nhiên của r_{ix} theo (3.14).

$$w_{is} = \begin{cases} \dfrac{1}{|Item(i,s)|} \displaystyle\sum_{x \in Item(i,s)} r_{ix} \ , nếu \ \ |Item(i,x)| \geq \theta \\ \dfrac{1}{\theta} \displaystyle\sum_{x \in Item(i,s)} r_{ix} \quad , \ nếu \ \ |Item(i,x)| < \theta \end{cases} \tag{3.14}$$

Dễ dàng nhận thấy $w_{is} \in F$, trong đó $F = \{1,2,..,g\}$. Chính vì vậy, ta có thể xem mỗi đặc trưng nội dung sản phẩm đóng vai trò như một sản phẩm phụ bổ sung vào tập sản phẩm. Ma trận đánh giá mở rộng theo hồ sơ người dùng được xác định theo (3.15). Trong đó, $p_x = c_s$ ($c_s \in C$) đóng vai trò như một sản phẩm phụ bổ sung vào ma trận đánh giá về phía sản phẩm.

$$r_{ix} = \begin{cases} r_{ix} \ \ , nếu\ x \in P \\ w_{is} \ \ , nếu\ s \in C \ \ (x = s) \end{cases} \tag{3.15}$$

3.3.1.2. Hợp nhất hồ sơ sản phẩm của lọc nội dung vào ma trận đánh giá

Gọi $U_x \subseteq U$ được xác đinh theo công thức (3.16) là tập người dùng $u_i \in U$ đã sử dụng sản phẩm $p_x \in P$.

$$U_x = \{u_i \in U \mid r_{ix} \neq 0 \ \ (p_x \in P)\} \tag{3.16}$$

Gọi User (x, q) là tập người dùng $u_i \in U_x$ có đặc trưng $t_q \in T$ được xác định theo công thức (3.17).

$$User(x,q) = \{u_i \in U_x \mid t_{iq} \neq 0 \ (p_x \in P, t_q \in T\} \quad (3.17)$$

Tác giả đề xuất phương pháp trích chọn đặc trưng nội dung người dùng có cùng mức độ đánh giá với giá trị đánh giá r_{ix} theo (3.18).

$$v_{qx} = \begin{cases} \frac{1}{|User(x,q)|} \sum_{i \in User(x,q)} r_{ix}, & nếu \ \ |User(x,q)| \geq \theta \\ \frac{1}{\theta} \sum_{i \in User(x,q)} r_{ix}, & nếu \ \ |User(x,q)| < \theta \end{cases} \quad (3.18)$$

Ma trận đánh giá mở rộng theo hồ sơ sản phẩm được xác định theo công thức (3.19). Trong đó, $u_i = t_q$ $(t_q \in T)$ đóng vai trò như một người dùng phụ bổ sung vào để mở rộng ma trận đánh giá về phía người dùng.

$$r_{ix} = \begin{cases} r_{ix} & ,nếu\ u_i \in U\ và\ r_{ix} \neq 0 \\ v_{qx} & ,nếu\ \ t_q \in T\ và\ v_{qx} \neq 0\ (u_i = t_q) \end{cases} \quad (3.19)$$

3.3.2. Mô hình học theo người dùng

Mô hình học kết hợp theo người dùng phát triển từ mô hình học theo người dùng cho lọc cộng tác đề xuất trong Mục 3.2.3.1.

Để hạn chế ảnh hưởng của vấn đề dữ liệu thưa, với mỗi người dùng $u_i \in U$ tác giả xây dựng tập sinh S_i được định nghĩa theo (3.20) để giám sát việc tính toán mức độ tương tự giữa các cặp người dùng. Trong đó, P_i được xác định theo (3.12), C_i được xác định theo (3.21).

$$S_i = \{u_j \in U: |P_i \cap P_j| \geq \theta_1 và\ |C_i \cap C_j| \geq \theta_2\} \quad (3.20)$$

$$C_i = \{c_s \in C: r_{is} \neq 0\} \quad (3.21)$$

Dựa vào S_i và độ tương quan Pearson, mức độ tương tự giữa các cặp người dùng của lọc cộng tác được xác định theo công thức (3.22), mức độ tương tự giữa các cặp người dùng của lọc nội dung được xác định theo công thức (3.23), mức độ tương tự giữa các cặp người dùng của lọc kết hợp được xác định theo công thức (3.24).

$$a_{ij} = \begin{cases} 0 & ,nếu\ u_j \notin S_i \\ \frac{\sum_{p_x \in P_i \cap P_j} (r_{ix} - \overline{r_i})(r_{jx} - \overline{r_j})}{\sqrt{\sum_{p_x \in P_i \cap P_j} (r_{ix} - \overline{r_i})^2} \sqrt{\sum_{p_x \in P_i \cap P_j} (r_{jx} - \overline{r_j})^2}} & ,nếu\ u_j \in S_i \end{cases} \quad (3.22)$$

$$b_{ij} = \begin{cases} 0 & ,nếu\ u_j \notin S_i \\ \frac{\sum_{c_s \in C_i \cap C_j} (r_{is} - \ddot{r_i})(r_{js} - \ddot{r_j})}{\sqrt{\sum_{c_s \in C_i \cap C_j} (r_{is} - \ddot{r_i})^2} \sqrt{\sum_{c_s \in C_i \cap C_j} (r_{js} - \ddot{r_j})^2}} & ,nếu\ u_j \in S_i \end{cases} \quad (3.23)$$

$$u_{ij} = \begin{cases} \frac{\sum_{p_x \in H_i \cap H_j} (r_{ix} - \overline{\overline{r_i}})(r_{jx} - \overline{\overline{r_j}})}{\sqrt{\sum_{p_x \in H_i \cap H_j} (r_{ix} - \overline{\overline{r_i}})^2} \sqrt{\sum_{p_x \in H_i \cap H_j} (r_{ix} - \overline{\overline{r_i}})^2}} \\ (nếu\ u_j \in S_i\ và\ a_{ij} \geq \alpha\ à\ b_{ij} \geq \alpha) \\ 0 \quad (trong\ các\ trường\ hợp\ khác) \end{cases} \quad (3.24)$$

Trong đó, P_i được xác định theo (3.12), C_i được xác định theo công thức (3.21); H_i, $\overline{r_i}$, $\ddot{r_i}$, $\overline{\overline{r_i}}$ được xác định tuần tự theo (3.25), (3.26), (3.27), (3.28).

$$H_i = P_i \cup C_i \tag{3.25}$$

$$\overline{r_i} = \frac{1}{|P_i \cap P_j|} \sum_{p_x \in P_i \cap P_j} r_{ix} \tag{3.26}$$

$$\ddot{r_i} = \frac{1}{|C_i \cap C_j|} \sum_{c_s \in C_i \cap C_j} r_{is} \tag{3.27}$$

$$\overline{\overline{r_i}} = \frac{1}{|H_i \cap H_j|} \sum_{p_x \in H_i \cap H_j} r_{ix} \tag{3.28}$$

Sau khi xác định được mức độ tương tự giữa các cặp người dùng, tác giả xây dựng tập láng giềng cho người dùng $u_i \in U$ theo công thức (3.29). Phương pháp dự đoán các sản phẩm mới $p_x \in P$ chưa được người dùng u_i biết đến được thực hiện theo công thức (3.30).

$$K_i = \{u_j \in S_i : u_{ij} > \alpha\} \tag{3.29}$$

$$r_{ix} = \overline{\overline{r_i}} + \frac{\sum_{u_j \in K_i} (r_{jx} - \overline{\overline{r_j}}) u_{ij}}{\sum_{u_j \in K_i} |u_{ij}|} \tag{3.30}$$

Những sản phẩm mới $p_x \in P$ có giá trị dự đoán r_{ix} theo (3.30) là những dự đoán tin cậy được bổ sung vào ma trận đánh giá mở rộng theo hồ sơ sản phẩm.

3.3.3. Mô hình học kết hợp theo sản phẩm

Mô hình học kết hợp theo sản phẩm phát triển từ mô hình học theo sản phẩm cho lọc cộng tác bằng phương pháp đồng huấn luyện đề xuất trong Mục 3.3.2.

Tương tự như người dùng, với mỗi sản phẩm $p_x \in P$ tác giả xây dựng tập S_x được định nghĩa theo công thức (3.31) để giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm. Trong đó, U_x được xác định theo công thức (3.16), T_x được xác định theo công thức (3.32).

$$S_x = \{p_y \in P : |U_x \cap U_y| \geq \gamma_1 \text{ và } |T_x \cap T_y| \geq \gamma_2\} \tag{3.31}$$

$$T_x = \{t_q \in T : r_{qx} \neq 0\} \tag{3.32}$$

Dựa vào S_x và độ tương quan Pearson, mức độ tương tự giữa các cặp sản phẩm của lọc cộng tác được xác định theo công thức (3.33), mức độ tương tự giữa các cặp sản phẩm của lọc nội dung được xác định theo công thức (3.34), mức độ tương tự giữa các cặp sản phẩm của lọc kết hợp được xác định theo công thức (3.35).

$$a_{xy} = \begin{cases} 0 & , nếu\ p_y \notin S_x \\ \dfrac{\sum_{u_i \in U_x \cap U_y} (r_{ix} - \overline{r_x})(r_{iy} - \overline{r_y})}{\sqrt{\sum_{u_i \in U_x \cap U_y} (r_{ix} - \overline{r_x})^2} \sqrt{\sum_{u_i \in U_x \cap U_y} (r_{iy} - \overline{r_y})^2}} & , nếu\ p_y \in S_x \end{cases} \tag{3.33}$$

$$b_{xy} = \begin{cases} 0 & , nếu\ p_y \notin S_x \\ \dfrac{\sum_{t_q \in T_x \cap T_y} (r_{qx} - \ddot{r_x})(r_{qy} - \ddot{r_y})}{\sqrt{\sum_{t_q \in T_x \cap T_y} (r_{qx} - \ddot{r_x})^2} \sqrt{\sum_{t_q \in T_x \cap T_y} (r_{qy} - \ddot{r_y})^2}} & , nếu\ p_y \in S_x \end{cases} \tag{3.34}$$

$$p_{xy} = \begin{cases} \dfrac{\sum_{u_i \in H_x \cap H_y} (r_{ix} - \overline{\overline{r_x}})(r_{iy} - \overline{\overline{r_y}})}{\sqrt{\sum_{u_i \in H_x \cap H_y} (r_{ix} - \overline{\overline{r_x}})^2} \sqrt{\sum_{u_i \in H_x \cap H_y} (r_{iy} - \overline{\overline{r_y}})^2}} \\ \quad (nếu\ \ p_y \in S_x\ và\ a_{xy} \geq \alpha\ và\ b_{xy} \geq \alpha) \\ 0\ \ (\ trong\ các\ trường\ hợp\ khác) \end{cases} \quad (3.35)$$

Trong đó, U_x được xác định theo công thức (3.16), T_x được xác định theo công thức (3.32), H_x, $\overline{r_x}$, $\ddot{r_x}$, $\overline{\overline{r_x}}$ được xác định theo công thức (3.36), (3.37), (3.38), (3.39), theo thứ tự.

$$H_x = U_x \cup T_x \quad (3.36)$$

$$\overline{r_x} = \frac{1}{|U_x \cap U_y|} \sum_{i \in U_x \cap U_y} r_{ix} \quad (3.37)$$

$$\ddot{r_x} = \frac{1}{|T_x \cap T_y|} \sum_{q \in T_x \cap T_y} r_{qx} \quad (3.38)$$

$$\overline{\overline{r_x}} = \frac{1}{|H_x \cap H_y|} \sum_{i \in H_x \cap H_y} r_{ix} \quad (3.39)$$

Sau khi xác định được mức độ tương tự giữa các cặp sản phẩm, tác giả xây dựng tập láng giềng cho sản phẩm $p_x \in P$ theo công thức (3.40). Phương pháp dự đoán mức độ phù hợp của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$ được thực hiện theo công thức (3.41).

$$K_x = \{y \in S_x:\ p_{xy} > \alpha\} \quad (3.40)$$

$$r_{ix} = \frac{\sum_{y \in K_x} p_{xy} r_{iy}}{\sum_{y \in K_x} |p_{xy}|} \quad (3.41)$$

Giá trị dự đoán r_{ix} theo (3.41) phản ánh mức độ phù hợp của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$ được bổ sung vào ma trận đánh giá mở rộng theo hồ sơ người dùng.

3.3.4. Mô hình đồng huấn luyện cho lọc kết hợp

Đầu vào:

- *Ma trận* $R = \{r_{ix}\}$ *được xác định theo công thức (3.9).*
- *Ma trận* $C = \{c_{xs}\}$ *được xác định theo công thức (3.10).*
- *Ma trận* $T = \{t_{iq}\}$ *được xác định theo công thức (3.11).*
- *Người dùng* $u_a \in U$ *là người dùng hiện thời cần được tư vấn.*
- *K là số lượng sản phẩm cần tư vấn cho người dùng hiện thời.*
- t_{max} *là số vòng lặp giới hạn.*

Đầu ra : *Danh sách K sản phẩm được tư vấn tới người dùng hiện thời* u_a.

Các bước tiến hành:

Begin

Bước 1(Khởi tạo):

$t \leftarrow 0$;*//khởi tạo số bước lặp ban đầu là* 0

$R^{(0)} = \left\{r_{ix}^{(0)} = r_{ix} : i = 1, 2, \dots, N; x = 1, 2, \dots, M\right\}$;

Bước 2 (Bước lặp):

Repeat

2.1. Tăng bước lặp : $t \leftarrow t+1$;

2.2. Huấn luyện kết hợp theo người dùng

a) Xác định trọng số các đặc trưng nội dung sản phẩm $w_{is}^{(t)}$ *tại vòng lặp thứ* t *theo công thức* (3.14).

b) Mở rộng ma trận đánh giá theo hồ sơ người dùng $r_{ix}^{(t)}$ *tại vòng lặp thứ t theo công thức* (3.15).

c) Xác định $S_i^{(t)}$ *theo công thức* (3.20).

d) Tính toán $u_{ij}^{(t)}$ *theo công thức* (3.24).

e) Xác định $K_i^{(t)}$ *theo công thức* (3.29).

f) Dự đoán giá trị $r_{ix}^{(t)}$ *theo công thức* (3.30).

2.3. Huấn luyện kết hợp theo sản phẩm

a) Xác định trọng số các đặc trưng nội dung người dùng $v_{qx}^{(t)}$ *tại vòng lặp thứ t theo công thức* (3.18).

b) *Mở rộng ma trận đánh giá theo hồ sơ sản phẩm* $r_{ix}^{(t)}$ *theo công thức* (3.19).

c) Xác định $S_x^{(t)}$ *theo công thức* (3.31).

d) Tính toán $p_{xy}^{(t)}$ *theo công thức* (3.35).

e) Xác định $K_x^{(t)}$ *theo công thức* (3.40).

f) Dự đoán giá trị $r_{ix}^{(t)}$ *theo công thức* (3.41).

Until (($r_{ix}^{(t)} = r_{ix}^{(t-1)}$) hoặc (t = t_{max}))

Bước 3 (sinh ra tư vấn):

<Sắp xếp các sản phẩm theo thứ tự giảm dần của $r_{ix}^{(t)}$*>;*

<Chọn K sản phẩm $p_x \in P$ *đầu tiên tư vấn cho người dùng* u_a*>;*

End.

Thuật toán 3.4. Thuật toán CoTraining–HybridFiltering

3.4. Thực nghiệm và kết quả

3.4.1. Thực nghiệm và kết quả của phương pháp lọc cộng tác bằng đồng huấn luyện

3.4.1.1.Dữ liệu thực nghiệm

Thuật toán lọc cộng tác được thực nghiệm trên 3 bộ dữ liệu: *MovieLens-100K* bao gồm 100.000 đánh giá của 943 người dùng cho 1682 phim; *MovieLens-1M* bao gồm 1000.000 đánh giá của 6000 người dùng cho 4000 phim; *MovieLens-10M* bao gồm 10.000.000 đánh giá của 72000 người dùng với 10.000 bộ phim.

3.4.1.2.Cài đặt thực nghiệm

- **Độ đo:** $MAE, RMSE$.

- **Phương pháp thực nghiệm:** Việc phân chia tập dữ liệu U thành 2 tập U_{train} và U_{test} được thực hiện như sau: Lần lượt chọn ngẫu nhiên 200, 400, và 600 người dùng trong tập MovieLens-100K làm dữ liệu huấn luyện, 200 người dùng được lựa chọn ngẫu nhiên trong số còn lại để làm tập kiểm tra. Chọn ngẫu nhiên 1000, 2000, và 3000 người dùng trong tập MovieLens-1M làm dữ liệu huấn luyện, 1000 người dùng được lựa chọn ngẫu nhiên trong số còn lại để làm tập kiểm tra. Chọn ngẫu nhiên 10000, 20000, và 40000 người dùng trong tập MovieLens-10M làm dữ liệu huấn luyện, 10000 người dùng được lựa chọn ngẫu nhiên trong số còn lại để làm tập kiểm tra. Việc thực nghiệm được thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.
- **Các phương pháp tư vấn được sử dụng để so sánh:** *UserBased, ItemBased, CoTraining-UserItem, CoTraining-ItemUser.*

3.4.1.3.Kết quả kiểm nghiệm

Bảng 3.11. Giá trị MAE, RMSE trên tập MovieLens-100K

Kích thước tập dữ liệu huấn luyện	Phương pháp	MAE			RMSE		
		Số đánh giá biết trước			Số đánh giá biết trước		
		5	10	20	5	10	20
200 người dùng	**UserBased**	0.732	0.711	0.645	0.934	0.908	0.824
	ItemBased	0.742	0.722	0.673	0.943	0.917	0.855
	CoTraining-UserItem	**0.621**	**0.594**	**0.512**	**0.789**	**0.754**	**0.651**
	CoTraining -ItemUser	**0.598**	**0.572**	**0.507**	**0.761**	**0.727**	**0.644**
400 người dùng	**UserBased**	0.694	0.675	0.644	0.885	0.862	0.822
	ItemBased	0.711	0.697	0.653	0.904	0.886	0.829
	CoTraining -UserItem	**0.615**	**0.615**	**0.587**	**0.782**	**0.781**	**0.746**
	CoTraining -ItemUser	**0.607**	**0.607**	**0.517**	**0.771**	**0.769**	**0.657**
600 người dùng	**UserBased**	0.693	0.686	0.686	0.885	0.876	0.876
	ItemBased	0.697	0.687	0.687	0.886	0.873	0.873
	CoTraining -UserItem	**0.548**	**0.519**	**0.511**	**0.696**	**0.659**	**0.649**
	CoTraining -ItemUser	**0.534**	**0.524**	**0.514**	**0.679**	**0.666**	**0.653**

Bảng 3.12. Giá trị MAE, RMSE trên tập MovieLens-1M

Kích thước tập dữ liệu huấn luyện	Phương pháp	MAE			RMSE		
		Số đánh giá biết trước			Số đánh giá biết trước		
		5	10	20	5	10	20
1000 người dùng	**UserBased**	0.792	0.779	0.764	0.960	0.945	0.927
	ItemBased	0.789	0.774	0.732	0.952	0.934	0.883
	CoTraining-UserItem	**0.764**	**0.752**	**0.716**	**0.922**	**0.906**	**0.864**
	CoTraining -ItemUser	**0.759**	**0.756**	**0.714**	**0.917**	**0.912**	**0.862**
2000 người dùng	**UserBased**	0.734	0.725	0.663	0.889	0.879	0.803
	ItemBased	0.731	0.739	0.657	0.883	0.892	0.792
	CoTraining -UserItem	**0.685**	**0.654**	**0.615**	**0.827**	**0.789**	**0.743**
	CoTraining -ItemUser	**0.667**	**0.647**	**0.607**	**0.805**	**0.779**	**0.733**
4000 người dùng	**UserBased**	0.713	0.688	0.686	0.865	0.835	0.832
	ItemBased	0.719	0.675	0.618	0.868	0.815	0.746
	CoTraining -UserItem	**0.684**	**0.642**	**0.597**	**0.825**	**0.774**	**0.720**
	CoTraining -ItemUser	**0.667**	**0.631**	**0.598**	**0.806**	**0.761**	**0.721**

Bảng 0.1. Giá trị MAE, RMSE trên tập MovieLens-10M

Kích thước tập dữ liệu huấn luyện	Phương pháp	MAE			RMSE		
		Số đánh giá biết trước			Số đánh giá biết trước		
		5	10	20	5	10	20
10000 người dùng	**UserBased**	0.763	0.724	0.716	0.924	0.878	0.868
	ItemBased	0.788	0.729	0.723	0.951	0.879	0.873
	CoTraining-UserItem	**0.712**	**0.694**	**0.647**	**0.859**	**0.837**	**0.781**
	CoTraining -ItemUser	**0.708**	**0.674**	**0.653**	**0.856**	**0.813**	**0.788**
20000 người dùng	**UserBased**	0.734	0.615	0.664	0.889	0.746	0.805
	ItemBased	0.746	0.618	0.672	0.901	0.746	0.810
	CoTraining -UserItem	**0.689**	**0.643**	**0.622**	**0.832**	**0.775**	**0.751**
	CoTraining -ItemUser	**0.681**	**0.667**	**0.619**	**0.822**	**0.802**	**0.747**
40000 người dùng	**UserBased**	0.796	0.766	0.684	0.965	0.929	0.829
	ItemBased	0.790	0.775	0.698	0.954	0.936	0.843
	CoTraining -UserItem	**0.688**	**0.669**	**0.616**	**0.831**	**0.807**	**0.743**
	CoTraining -ItemUser	**0.679**	**0.654**	**0.642**	**0.820**	**0.789**	**0.774**

Kết quả kiểm nghiệm đưa ra trong Bảng 3.11, Bảng 3.12, và Bảng 3.13 cho thấy sai số $MAE, RMSE$ của cả hai phương pháp lọc cộng tác bằng đồng huấn luyện CoTraining-UserItem và CoTraining-ItemUser đều nhỏ hơn UserBased và ItemBased truyền thống trên mọi kích thước dữ liệu huấn luyện và số lượng đánh giá cho trước của người dùng. Điều đó có thể khẳng định phương pháp đề xuất cải thiện đáng kể chất lượng dự đoán cho lọc cộng tác, đặc biệt trong trường hợp dữ liệu thưa.

3.4.2. Thực nghiệm và kết quả của phương pháp lọc kết hợp bằng đồng huấn luyện

3.4.2.1.Dữ liệu thực nghiệm

Tác giả sử dụng tập dữ liêu MovieLens 1M để tiến hành thực nghiệm cho phương pháp đề xuất. Tập dữ liệu MovieLens 1M gồm 1MB đánh giá của 6000 người dùng cho 4000 phim.

3.4.2.2.Cài đặt thực nghiệm

- **Độ đo:** $MAE, RMSE$.
- **Phương pháp thực nghiệm:** việc phân chia tập dữ liệu U thành 2 tập U_{train} và U_{test} được thực hiện như sau: Lấy ngẫu nhiên 4000 người dùng trong tập MovieLens làm dữ liệu huấn luyện. Chọn ngẫu nhiên 1000 người dùng trong số còn lại để làm 4 tập dữ liệu kiểm tra (test1.inp, test2.inp, test3.inp, test3.inp). Đối với mỗi tập dữ liệu kiểm tra, tác giả thực hiện loại bỏ ngẫu nhiên các đánh giá sao cho số các đánh giá biết trước của mỗi người dùng đối với sản phẩm chỉ còn lại là 5, 10, 15 và 20 đánh giá.
- **Các phương pháp tư vấn được sử dụng để so sánh:** *CF-UserBased, CF-ItemBased, CBF-UserBased, CBF-ItemBased, Hybrid-UserBased, Hybrid-ItemBased, CoTraining-HybridFiltering.*

3.4.2.3.Kết quả kiểm nghiệm

Kết quả trong Bảng 3.14 cho thấy phương pháp CoTraining- HybridFiltering cho lại giá trị MAE, RMSE thấp nhất ở tất cả các mức độ thưa thớt dữ liệu khác nhau. Điều này có thể khẳng định phương pháp xác định độ tương tự dựa trên tập không thưa đối với người dùng và sản phẩm là hoàn toàn tin cậy. Phương pháp đồng huấn luyện cho lọc kết hợp đề xuất cho phép chuyển giao kết quả

dự đoán giữa quá trình học kết hợp theo người dùng và học kết hợp theo sản phẩm để hạn chế hiệu quả vấn đề dữ liệu thưa của các phương pháp lọc.

Phương pháp	MAE				RMSE			
	Số lượng đánh giá biết trước				Số lượng đánh giá biết trước			
	5	10	15	20	5	10	15	20
CBF-UserBased	0.865	0.859	0.855	0.835	1.049	1.042	1.029	1.013
CBF-ItemBased	0.894	0.883	0.875	0.845	1.085	1.071	1.054	1.025
CF-UserBased	0.824	0.817	0.821	0.813	0.999	0.992	0.988	0.986
CF-ItemBased	0.846	0.841	0.836	0.815	1.021	1.015	0.998	0.984
Hybrid-UserBased	**0.793**	**0.792**	**0.791**	**0.702**	**0.957**	**0.956**	**0.946**	**0.922**
Hybrid-ItemBased	**0.798**	**0.788**	**0.782**	**0.695**	**0.963**	**0.952**	**0.935**	**0.928**
CoTraining-HybridFiltering	**0.672**	**0.629**	**0.617**	**0.585**	**0.811**	**0.759**	**0.738**	**0.707**

Để đánh giá về mức độ ảnh hưởng của việc tích hợp thêm đặc trưng nội dung vào phương pháp đồng huấn luyện cho lọc kết hợp so với phương pháp đồng huấn luyện cho lọc cộng tác, ta quan sát kết quả kiểm nghiệm của phương pháp *CoTraining- HybridFiltering* trong bảng 3.14 và *CoTraining -UserItem* trong bảng 3.12 trong trường hợp sử dụng cùng 4000 người dùng làm dữ liệu huấn luyện. Kết quả MAE của *CoTraining –UserItem* là 0.684, 0.642, 0.597, trong khi đó MAE của CoTraining- HybridFiltering là 0.672, 0.629, 0.617, 0.585 với lần lượt các mức độ thưa thớt 5, 10, 20 đánh giá biết trước. Nhận định tương tự khi so sánh giá trị RMSE của hai phương pháp này. Điều đó chứng tỏ độ chính xác dự đoán đánh giá của phương pháp lọc kết hợp được cải thiện khi tích hợp thêm đặc trưng nội dung vào quá trình đồng huấn luyện hơn so với phương pháp lọc cộng tác bằng đồng huấn luyện.

3.5. Kết luận chương 3

Chương này đã trình bày kết quả nghiên cứu của luận án về đề xuất một phương pháp lọc kết hợp mới giữa lọc cộng tác và lọc nội dung. Mô hình kết hợp giữa lọc cộng tác và lọc nội dung được trình bày trong chương này thực hiện dựa trên việc hợp nhất biểu diễn các giá trị đặc trưng nội dung vào lọc cộng tác. Lọc kết hợp bằng phương pháp đồng huấn luyện đề xuất phát triển từ phương pháp lọc cộng tác bằng phương pháp đồng huấn luyện, đây là một phương pháp thuộc hướng tiếp cận học bán giám sát cho bài toán phân lớp. Trong đó, quá trình huấn luyện theo người dùng bổ sung thêm một số nhãn phân loại chắc chắn cho quá trình huấn luyện theo sản phẩm. Ngược lại, quá trình huấn luyện theo sản phẩm bổ sung thêm các nhãn phân loại chắc chắn cho quá trình huấn luyện theo người dùng. Hai quá trình huấn luyện thực hiện đồng thời cho phép bổ sung các nhãn phân loại tin cậy theo mỗi bước thực hiện, nhờ vậy cải thiện độ chính xác dự đoán đánh giá và tư vấn sản phẩm phù hợp cho người dùng. Kết quả thực nghiệm trên bộ dữ liệu thực về phim cho thấy, phương pháp đề xuất cho lại kết quả dự đoán khá tốt, đặc biệt trong trường hợp dữ liệu thưa.

KẾT LUẬN

I. Kết quả đạt được của luận án

1. Về mặt lý thuyết, luận án tổng kết những nghiên cứu cơ bản và mở rộng về hệ tư vấn theo các hướng tiếp cận khác nhau, kèm theo những vấn đề cần tiếp tục nghiên cứu và xu hướng. Trên cơ sở những kiến thức nền tảng, tác giả tập trung nghiên cứu nâng cao kết quả dự đoán sản phẩm cho người dùng trong trường hợp dữ liệu thưa, cũng như trong trường hợp có cả dữ liệu sở thích

người dùng, thông tin nội dung người dùng, thông tin nội dung sản phẩm và thông tin ngữ cảnh sử dụng sản phẩm của người dùng. Kết quả luận án đưa ra 2 đề xuất chính: 1) Đề xuất một phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh [C1][C3][C7][C4][J2]; 2) Đề xuất một phương pháp lọc kết hợp bằng phương pháp đồng huấn luyện [C2][C5][C6][J1].

2. Về mặt thực tiễn, kết quả của luận án đã được thực nghiệm trên các bộ dữ liệu thực trong các kịch bản khác nhau, kết quả thực nghiệm của phương pháp đề xuất được đánh giá là có độ chính xác tốt hơn các phương pháp cơ sở trong đa số trường hợp, đồng thời đơn giản trong cài đặt để triển khai các hệ tư vấn thực tế. Đây là sở cứ cho thấy có thể áp dụng kết quả nghiên cứu của đề tài trong việc triển khai các hệ thống tư vấn thông tin cá nhân hóa tới người dùng ở đa dạng các lĩnh vực.

II. Hạn chế và hướng phát triển của luận án

1. Hạn chế

   Một số hạn chế nhất định chưa được giải quyết trong các đề xuất nêu ra bởi luận án, đó là:

   - Vấn đề sở thích của người dùng với sản phẩm thay đổi cập nhật thường xuyên theo thời gian.
   - Vấn đề người dùng mới tham gia vào hệ thống tư vấn.

2. Hướng phát triển

   - Nghiên cứu phát triển mô hình học máy mới cho hệ tư vấn theo hướng kết hợp thông tin nội dung về đặc trưng sản phẩm và người dùng trong hệ tư vấn theo ngữ cảnh.
   - Nghiên cứu phát triển phương pháp đồng huấn luyện cho lọc cộng tác và lọc kết hợp theo hướng mở rộng nhiều cơ chế quan sát dữ liệu phù hợp với từng bộ dữ liệu thực tế. Đồng thời xem xét tích hợp những mô hình phân lớp tiên tiến để học dữ liệu.
   - Nghiên cứu giải quyết vấn đề người dùng mới, sở thích của người dùng với sản phẩm thay đổi theo thời gian.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ CỦA NGHIÊN CỨU SINH

C1. Do Thi Lien, Nguyen Duy Phuong: Collaborative filtering with a graph-based similarity measure. 2014 International Conference on Computing, Management and Telecommunications, ComManTel 2014, pp 251–256 (2014).

C2. Tran Nhat Quang, Do Thi Lien, and Nguyen Duy Phuong: Collaborative Filtering by Co-Training Method. Knowledge and Systems Engineering 2014 Sixth International Conference on Knowledge and Systems Engineering, pp 273-285 (2014).

C3. Do Thi Lien, Nguyen Xuan Anh, Nguyen Duy Phuong: A Graph Model For Hybrid Recommender System. Knowledge and Systems Engineering 2015 Seventh International Conference on Knowledge and Systems Engineering, pp 138-143 (2015).

C4. Đỗ Thị Liên, Nguyễn Xuân Anh, Nguyễn Duy Phương, Từ Minh Phương: Một mô hình đồ thị cho hệ tư vấn lai. Fair'8 - Nghiên Cứu Cơ Bản Và Ứng Dụng Công Nghệ Thông Tin, trang 430-443 (2015).

C5. Do Thi Lien, Nguyen Duy Phuong: A Semi-supervised Learning Method for Hybrid

Filtering. ICTA International Conference on Advances in Information and Communication Technology. 538, pp 94-103 (2016).

C6. Đỗ Thị Liên, Nguyễn Duy Phương: Một Phương Pháp Học Bán Giám Sát Cho Lọc Kết Hợp. Fair'9 - Nghiên Cứu Cơ Bản Và Ứng Dụng Công Nghệ Thông Tin, trang 423-434 (2016).

J1. Đỗ Thị Liên, Nguyễn Duy Phương, Từ Minh Phương: Hợp nhất lọc cộng tác và lọc nội dung bằng phương pháp học bán giám sát. Chuyên san các công trình nghiên cứu phát triển CNTT & TT. Tập V-2, số 18 (38), trang 1-11 (2017).

C7. Đỗ Thị Liên, Nguyễn Duy Phương: Một phương pháp tư vấn cộng tác theo ngữ cảnh. Fair 11 - Nghiên Cứu Cơ Bản Và Ứng Dụng Công Nghệ Thông Tin, trang 319-329 (2018).

J2. Tu Minh Phuong, Do Thi Lien, Nguyen Duy Phuong: Graph-based Context-Aware Collaborative Filtering. Expert Systems with Applications. 126, pp 9–19 (2019).