HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐỖ THỊ LIÊN

PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP XÂY DỰNG HỆ TƯ VẤN

LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2020

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

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. GS.TS. TỪ MINH PHƯƠNG
2. TS. NGUYỄN DUY PHƯƠNG

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố trong các công trình nào khác.

Tác giả

Đỗ Thị Liên

LỜI CẢM ƠN

Trong quá trình thực hiện đề tài "Phát triển một số phương pháp xây dựng hệ tư vấn", tôi đã nhận được rất nhiều sự giúp đỡ, tạo điều kiện của tập thể giáo viên hướng dẫn, nhà trường, đồng nghiệp, các nhà khoa học và gia đình. Tôi xin bày tỏ lòng cảm ơn chân thành về sự giúp đỡ đó.

Trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc tới tập thể giáo viên hướng dẫn GS.TS Từ Minh Phương và TS Nguyễn Duy Phương - những người Thầy trực tiếp hướng dẫn và chỉ bảo cho tôi hoàn thành luận án này. Cảm ơn hai Thầy rất nhiều vì sự hướng dẫn tận tình, nghiêm túc và khoa học.

Tôi xin trân trọng cảm ơn Hội đồng Khoa học, Hội đồng Tiến sỹ, Khoa Quốc tế và Đào tạo sau đại học của Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện thuận lợi cho tôi được thực hiện và hoàn thành chương trình nghiên cứu của mình.

Tôi xin cảm ơn tập thể Lãnh đạo, cán bộ, giảng viên khoa Công nghệ thông tin, khoa Đa phương tiện - Học viện Công nghệ Bưu chính Viễn thông đã cổ vũ động viên tôi trong suốt quá trình nghiên cứu.

Tôi cảm ơn tất cả những người bạn của tôi, những người luôn chia sẻ, cổ vũ tôi trong lúc khó khăn và tôi luôn ghi nhớ điều đó.

Cuối cùng, tôi xin bày tỏ lòng biết ơn chân thành đối với gia đình đã luôn động viên, ủng hộ, cổ vũ và tạo mọi điều kiện giúp đỡ tôi.

MỤC LỤC

Trang

DANH MỤC CÁC CHỮ VIẾT TẮT

KÝ HIỆU	DIỄN GIẢI	
	TIẾNG ANH	TIẾNG VIỆT
RS	Recommender System / Recommendation System	Hệ tư vấn
CARS	Context-Aware Recommender System	Hệ tư vấn theo ngữ cảnh
CF	Collaborative Filtering	Lọc cộng tác
CBF	Content-Based Filtering	Lọc theo nội dung
HF	Hybrid Filtering	Lọc kết hợp
IR	Information Retrieval	Truy vấn thông tin
MAE	Mean Absolute Error	Trung bình giá trị tuyệt đối lỗi
MAP	Mean Average Precision	Độ chính xác trung bình tuyệt đối
AP	Average Precision	Độ chính xác trung bình
RMSE	Root Mean Square Error	Trung bình lỗi lấy căn
KNN	K-Nearest Neighbor	K láng giềng gần nhất
SDP	Sparsity Data Problem	Vấn đề dữ liệu thưa
User-Based k-NN	User-Based k Neareast Neighbor	Phương pháp K láng giềng gần nhất dựa vào người dùng
Item-Based k-NN	Item-Based k Neareast Neighbor	Phương pháp K láng giềng gần nhất dựa vào sản phẩm
TF/IDF	Term Frequency / Inverse Document Frequency	Phép đo tần suất kết hợp với tần suất xuất hiện ngược
MD matrix	Multi-dimensional matrix	Ma trận đánh giá đa chiều

DANH MỤC HÌNH VẼ

Trang

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC THUẬT TOÁN

Trang

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. Thông thường khi cần tìm thông tin về một sản phẩm nào đó, giải pháp được hầu hết người dùng sử dụng là đưa câu hỏi vào máy tìm kiếm (Search engine) thay vì tìm đến những trang Web hoặc diễn đàn chuyên ngành. Máy tìm kiếm tiến hành tìm kiếm thông tin dựa trên các từ khóa (Keyword) được người dùng gõ vào và trả về một danh mục của các trang Web có chứa từ khóa mà nó tìm được. Do vậy việc sử dụng máy tìm kiếm sẽ hiệu quả khi người dùng biết họ thực sự muốn tìm cái gì. Trong trường hợp khi người dùng không xác định được chính xác cái mình muốn tìm thì yêu cầu về lọc thông tin một cách có hiệu quả và tin cậy là rất cần thiết. Để đáp ứng nhu cầu này, các hệ thống tư vấn đã ra đời, ví dụ một số hệ tư vấn đã được thương mại hóa và triển khai thành công, tiêu biểu là hệ tư vấn của các hãng Amazon, eBay, Netflix, Youtube …

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). Cụ thể, hệ tư vấn cung cấp một giải pháp giảm tải thông tin bằng cách đưa ra dự đoán đánh giá mức độ thích của người dùng với sản phẩm mới và cung cấp một danh sách ngắn các sản phẩm (trang web, bản tin, phim, video…) mà nhiều khả năng người dùng sẽ quan tâm [1]. Trên thực tế, hệ tư vấn không chỉ hướng đến vấn đề giảm tải thông tin cho mỗi người dùng mà nó còn là yếu tố quyết định đến thành công của các hệ thống thương mại điện tử [1][2].

Hệ tư vấn đang ngày càng trở thành một lĩnh vực nghiên cứu quan trọng từ sau khi xuất hiện bài báo đầu tiên về lọc cộng tác vào giữa những năm 90 [3]. Đã có rất nhiều công việc được thực hiện cả trong ngành công nghiệp và nghiên cứu hàn lâm

để phát triển các hệ tư vấn trong hơn thập kỷ qua. Về cơ bản hệ tư vấn được chia thành hai hướng tiếp cận chính [4][5] 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 đó, cách tiếp cận truyền thống sử dụng một trong ba phương pháp lọc tin chính (*Lọc cộng tác, lọc theo nội dung* và *lọc kết hợp*) lên ba loại thông tin đầu vào phổ biến (Thông tin người dùng, thông tin sản phẩm và phản hồi của người dùng về sản phẩm). Cách tiếp cận mở rộng đề cập ở đây được biết đến với một số hướng như: Hệ tư vấn theo ngữ cảnh (Context-aware Recommender Systems) [6][7], hệ tư vấn dựa trên mạng xã hội (Social-based Recommender Systems) [8], hệ tư vấn dựa trên mối quan tâm (Attention-based Recommender Systems) [9] hoặc phát triển các phương pháp lọc kết hợp. Theo đó, bên cạnh các loại thông tin điển hình của hệ tư vấn theo cách truyền thống, cách tiếp cận mở rộng này cho phép tích hợp thêm đa dạng các nguồn thông tin đầu vào (Thông tin ngữ cảnh, liên kết từ mạng xã hội, mối quan tâm …) nhằm cải thiện chất lượng của hệ tư vấn thực tế.

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 [1][4][5][10], 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

Đóng góp thứ nhất của luận án là 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]. Những đóng góp cụ thể của luận án bao gồm:

- Đề xuất độ đo tương tự giữa các cặp người dùng hoặc giữa các cặp sản phẩm cho lọc cộng tác dựa trên mô hình đồ thị. Độ đo tương tự đề xuất cho phép khai thác các mối quan hệ trực tiếp và bắc cầu giữa các đỉnh người dùng hoặc giữa các đỉnh sản phẩm trên đồ thị vào quá trình dự đoán và tư vấn, điều này giúp hạn chế ảnh hưởng của vấn đề thưa dữ liệu đánh giá. Đây chính là ưu điểm nổi bật của độ đo tương tự đề xuất so với các độ đo tương tự dựa vào bộ nhớ trước đây trong việc giải quyết bài toán lọc cộng tác theo bộ nhớ cho hệ tư vấn truyền thống.
- Phát huy những điểm mạnh của độ đo tương tự đề xuất nêu trên bằng việc mở rộng phạm vi áp dụng nó cho phát triển hệ tư vấn cộng tác theo ngữ cảnh. Phương pháp lọc cộng tác theo ngữ cảnh đề xuất ngoài việc giải quyết khá tốt vấn đề dữ liệu thưa, còn cho phép tích hợp đầy đủ thông tin ngữ cảnh vào quá trình dự đoán sản phẩm tới người dùng. Khi đó, các sản phẩm mới tư vấn cho người dùng sẽ được cá nhân hóa tốt hơn theo từng ngữ cảnh cụ thể. Phương pháp đề xuất được đánh giá là đơn giản trong cài đặt để triển khai cho các hệ tư vấn theo ngữ cảnh thực tế.

- Kết quả thực nghiệm và đánh giá trên một số bộ dữ liệu thực cho thấy phương pháp đề xuất cải thiện đáng kể chất lượng tư vấn.

Đóng góp thứ hai của luận án là đề xuất một phương pháp lọc kết hợp bằng phương pháp đồng huấn luyện [C2][C5][C6][J1]. Những đóng góp cụ thể của luận án bao gồm:

- Đề xuất phương pháp lọc cộng tác bằng phương pháp đồng huấn luyện. Phương pháp lọc cộng tác đề xuất cho phép giải quyết vấn đề thưa của dữ liệu đánh giá.
- Hợp nhất biểu diễn các giá trị đặc trưng nội dung vào lọc cộng tác. Việc hợp nhất biểu diễn này được tiếp cận theo 2 cơ chế quan sát dữ liệu: 1) Quan sát theo người dùng cho phép hợp nhất hồ sơ người dùng của lọc nội dung vào ma trận đánh giá; 2) Quan sát theo sản phẩm cho phép hợp nhất hồ sơ sản phẩm của lọc nội dung vào ma trận đánh giá.
- Sử dụng hợp nhất biểu diễn các giá trị đặc trưng nội dung vào lọc cộng tác để xây dựng phương pháp dự đoán cho lọc kết hợp bằng đồng huấn luyện. Phương pháp lọc kết hợp đề xuất phát triển từ phương pháp lọc cộng tác bằng đồng huấn luyện cho phép giải quyết vấn đề dữ liệu thưa, đồng thời tích hợp đầy đủ thông tin người dùng, sản phẩm và đánh giá của người dùng với sản phẩm vào quá trình dự đoán đánh giá.
- Kết quả thực nghiệm và đánh giá trên các bộ dữ liệu thực về phim cho thấy phương pháp đề xuất cải thiện đáng kể chất lượng tư vấn.

4. Bố cục của luận án

Luận án được tổ chức thành ba chương, trong đó :

Chương 1. Tổng quan về hệ tư vấn

Nội dung chính của chương trình bày những nghiên cứu cơ bản về hệ tư vấn, các phương pháp tiếp cận phổ biến trong xây dựng hệ tư vấn kèm theo những vấn

đề cần tiếp tục nghiên cứu và xu hướng. Trên cơ sở đó xác định rõ hướng nghiên cứu của đề tài.

Chương 2. Phát triển phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh

Trình bày phương pháp hạn chế ảnh hưởng vấn đề dữ liệu thưa của lọc cộng tác dựa trên mô hình đồ thị, mở rộng cho phát triển hệ tư vấn cộng tác theo ngữ cảnh. Nội dung trình bày trong chương được tổng hợp từ kết quả nghiên cứu đã công bố trong [C1][C3][C7][C4][J2].

Chương 3. Phát triển phương pháp lọc kết hợp bằng đồng huấn luyện

Trình bày phương pháp kết hợp giữa lọc cộng tác và lọc nội dung bằng đồng huấn luyện. Nội dung trình bày trong chương được tổng hợp từ kết quả nghiên cứu được công bố trong [C2][C5][C6][J1].

Cuối cùng là một số kết luận và hướng nghiên cứu tiếp theo.

CHƯƠNG 1: TỔNG QUAN VỀ HỆ TƯ VẤN

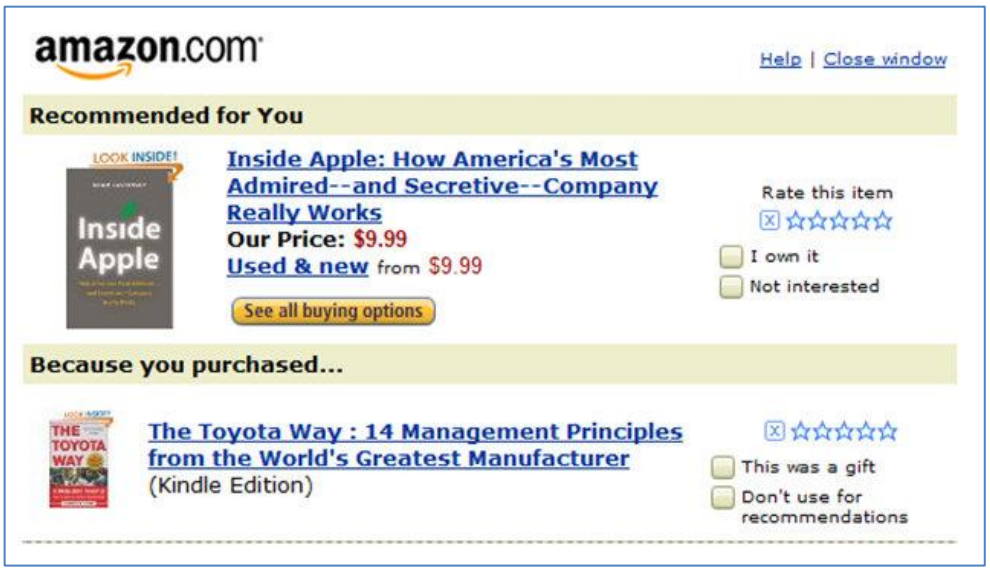
Mục tiêu chính của chương này trình bày những vấn đề tổng quan về hệ tư vấn, các phương pháp tiếp cận phổ biến trong xây dựng hệ tư vấn, phân tích rõ những hạn chế tồn tại của mỗi phương pháp và xu hướng phát triển hệ tư vấn trong những năm gần đây. Trên cơ sở những nghiên cứu cơ bản, xác định rõ hướng nghiên cứu cụ thể của đề tài. Những kết quả nghiên cứu của đề tài sẽ được trình bày trong các chương tiếp theo của luận án.

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 [1].

Theo Ricci và cộng sự [7], 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...

Ví dụ giao diện hệ tư vấn sách của Amazon:



Hình 1.1. Giao diện hệ tư vấn sách của Amazon

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 [11], điển hình như :

- Thương mại điện tử: Gợi ý những sản phẩm hoặc dịch vụ mua bán trực tuyến. Ví dụ hệ thống của Amazon – amazon.com, Ebay – ebay.com.
- Giáo dục: Gợi ý nguồn tài nguyên học tập như sách, bài báo, khóa học, địa chỉ Web,… cho người học. Ví dụ hệ thống của Foxtrot, InfoFinder.
- Giải trí: Gợi ý bài hát cho người nghe (Ví dụ hệ thống của LastFM - www.last.fm), gợi ý phim ảnh (Ví dụ hệ thống của Netflix, MovieLens, EachMovie), gợi ý các video clip (Ví dụ hệ thống của YouTube - www.youtube.com).
- Du lịch: Gợi ý điểm đến, hoạt động du lịch. Ví dụ hệ thống của Dietorecs, LifestyleFinder.
- Chăm sóc sức khỏe: Gợi ý sản phẩm y tế. Ví dụ hệ thống mạng xã hội sức khỏe – www.patientslikeme.com.
- Truyền thông xã hội: Gợi ý các hoạt động xã hội. Ví dụ hệ thống của Facebook, Twitter, LinkedIn.
- Ăn uống: Gợi ý nhà hàng, địa điểm ăn uống. Ví dụ hệ thống của Adaptive Place Advisor, Polylens, Pocket restaurant finder.

Bên cạnh đó, hệ tư vấn đã và đang được các nhà khoa học, các tổ chức, doanh nghiệp rất quan tâm nghiên cứu ứng dụng hệ tư vấn cho đa dạng các lớp bài toán ở các lĩnh vực khác nhau của cuộc sống.

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). Ví dụ $u_i \in U$ là một người dùng thì các đặc trưng nội dung biểu diễn người dùng u_i có thể là T ={*giới tính, độ tuổi, nghề nghiệp, trình độ,…*}.

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. Ví dụ $p_x \in P$ là một phim thì các đặc trưng nội dung biểu diễn phim p_x có thể là C = {*thể loại phim, nước sản xuất, hãng phim, diễn viên, đạo diễn,…*}.

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

Giá trị r_{ix} thể hiện đánh giá của người dùng $u_i \in U$ cho một số sản phẩm $p_x \in P$. Thông thường giá trị r_{ix} nhận một giá trị thuộc miền $F = \{1, 2, \dots, g\}$ được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Những giá trị $r_{ix} = 0$ được hiểu là người dùng $u_i \in U$ chưa biết đến hoặc không đánh giá sản phẩm $p_x \in P$, những ô điền ký tự "?" là giá trị cần hệ tư vấn đưa ra dự đoán đánh giá. Tiếp đến, ta ký hiệu $P_i \subseteq P$ là tập các sản phẩm $p_x \in P$ được đánh giá bởi người dùng $u_i \in U$ và $u_a \in U$

được gọi 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:

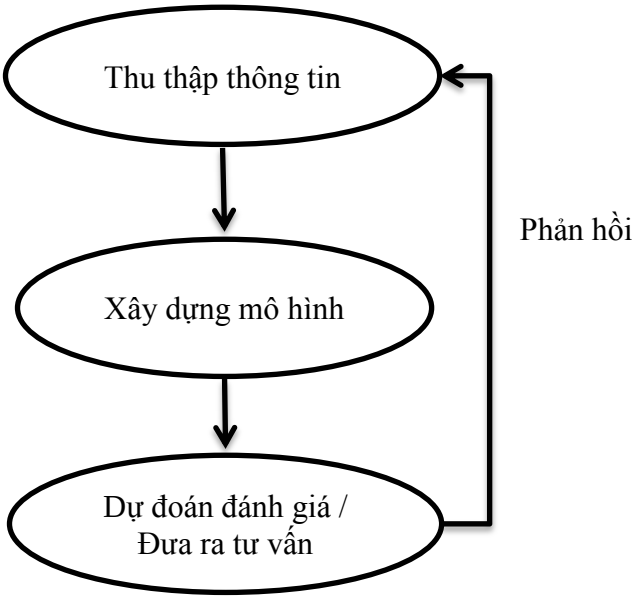
(1) Dự đoán đánh giá của người dùng u_a với các sản phẩm chưa có đánh giá trước đó.

(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.

Việc giải quyết bài toán tư vấn sẽ được thực hiện theo qui trình xây dựng hệ tư vấn trong mục 1.4 sau đây.

1.4. Qui trình xây dựng hệ tư vấn

Qui trình tổng quát để giải quyết bài toán tư vấn [12] thông thường gồm có 3 giai đoạn chính được thể hiện trong Hình 1.3 sau.



Hình 1.3. Qui trình xây dựng hệ tư vấn

Giai đoạn 1: Thu thập thông tin

Ba loại thông tin chính thường được thu thập cho hệ tư vấn, gồm có:

- Người dùng (User) biểu diễn thông qua các đặc trưng là thông tin cá nhân. Thông qua biểu diễn này, hệ thống cho phép xây dựng hồ sơ người dùng

(user's profile) nhằm lưu trữ lại dấu vết các đặc trưng nội dung sản phẩm đã từng được sử dụng bởi người dùng.

- Sản phẩm (Item) biểu diễn thông qua các đặc trưng là thông tin về sản phẩm. Thông qua biểu diễn này, hệ thống cho phép xây dựng hồ sơ sản phẩm (item's profile) nhằm lưu trữ lại dấu vết các đặc trưng người dùng đã từng sử dụng sản phẩm.
- Phản hồi của người dùng với sản phẩm (Feedback), biểu diễn thông qua các giá trị đánh giá của người dùng với sản phẩm.

Giai đoạn 2: Xây dựng mô hình

Giai đoạn xây dựng mô hình tư vấn có thể thực hiện bằng nhiều hướng tiếp cận khác nhau nhằm so sánh, đánh giá mối liên hệ giữa các thông tin thu thập được ở giai đoạn 1. Một số hướng tiếp cận điển hình được biết đến như: dựa vào kinh nghiệm (heuristics), học máy, lý thuyết xấp xỉ,...[1]. Mỗi hướng tiếp cận sẽ khai thác thông tin đầu vào theo những cách khác nhau hình thành những phương pháp tư vấn khác nhau. Chi tiết về các phương pháp tư vấn này sẽ được trình bày cụ thể trong mục 1.5 của luận án.

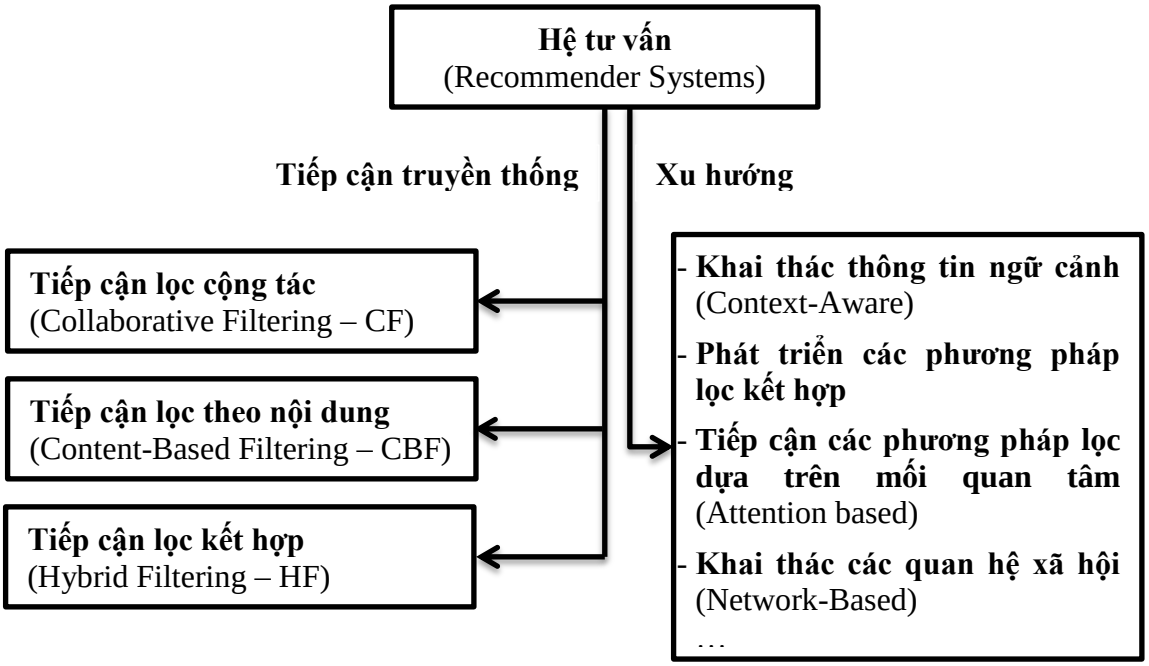
Giai đoạn 3: Dự đoán đánh giá / Đưa ra tư vấn

Dữ liệu đầu ra của giai đoạn 2 sẽ được dùng để dự đoán các đánh giá của người dùng với các sản phẩm chưa có đánh giá trước đó và chọn ra K sản phẩm mới phù hợp nhất đối với người dùng hiện thời để gợi ý cho họ.

1.5. Các hướng tiếp cận xây dựng hệ tư vấn

Có nhiều đề xuất khác nhau để giải quyết bài toán tư vấn theo "Qui trình xây dựng hệ tư vấn". Tuy nhiên về cơ bản thì hệ tư vấn được chia thành hai hướng tiếp cận tùy vào việc lựa chọn loại thông tin, mô hình học và dự đoán sản phẩm mới cho người dùng [4][5], đó là: 1) Hệ tư vấn với cách tiếp cận truyền thống; 2) Hệ tư vấn mở rộng cách tiếp cận truyền thống. Trong đó:

- Cách tiếp cận truyền thống khai thác 3 loại thông tin đầu vào gồm người dùng, sản phẩm và phản hồi của người dùng về sản phẩm. Dựa vào cách xác định dự đoán đánh giá cho các sản phẩm đối với người dùng, hệ tư vấn thường được chia thành ba loại: Tư vấn dựa vào phương pháp *lọc cộng tác* (Collaborative Filtering Recommendation), tư vấn dựa vào phương pháp *lọc theo nội dung* (Content-Based Filtering Recommendation) và tư vấn dựa vào phương pháp *lọc kết hợp* (Hybrid Filtering Recommendation) [1][5][10].
- Cách tiếp cận mở rộng từ hệ tư vấn truyền thống cho phép tích hợp thêm các nguồn thông tin khác (Ngữ cảnh, thông tin trong mạng xã hội,...) hoặc cải tiến các phương pháp lọc tin truyền thống trong hệ tư vấn (Các phương pháp lọc kết hợp, các phương pháp lọc dựa trên mối quan tâm…). Từ đây hệ tư vấn được chia thành một số loại điển hình: Tư vấn dựa vào ngữ cảnh (Context-Aware Recommendation), tư vấn dựa vào mạng xã hội (Social Network-Based Recommendation), tư vấn dựa vào các phương pháp lọc kết hợp (Hybrid Filtering Recommendation), tiếp cận các phương pháp lọc dựa trên mối quan tâm (Attention-based Recommendation)… [1][4][9].

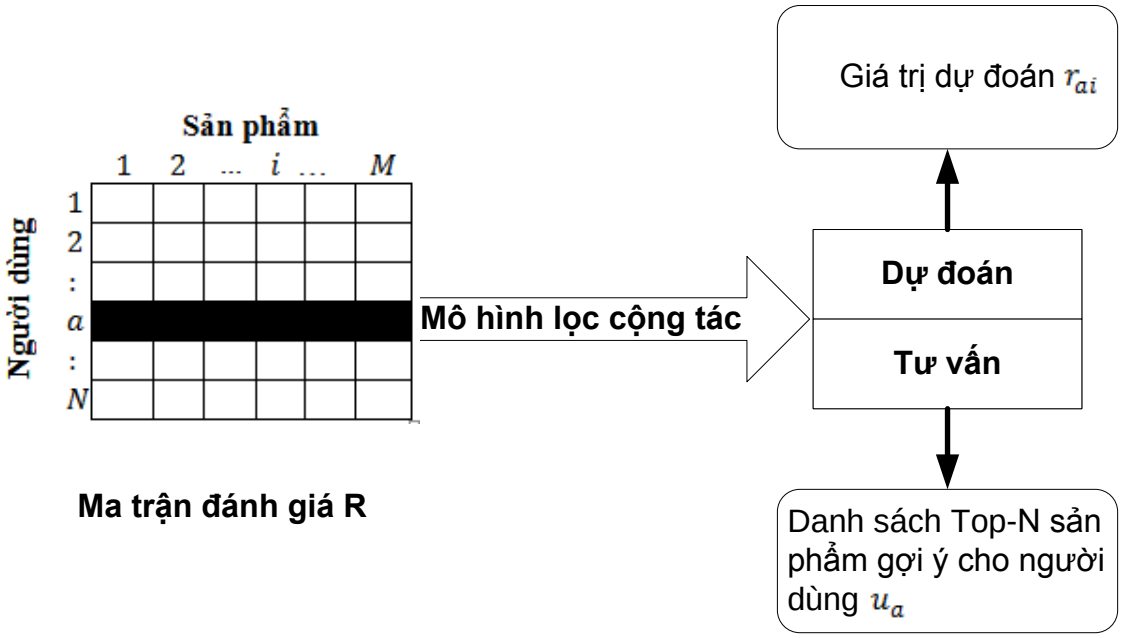


Hình 1.4. Các hướng tiếp cận truyền thống và xu hướng hiện nay của hệ tư vấn

Nội dung dưới đây sẽ trình bày chi tiết về các phương pháp này.

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 [1][10]. Như vậy thông tin đầu vào của hệ tư vấn dựa vào phương pháp lọc cộng tác chính là các phản hồi của người dùng về các sản phẩm trong hệ thống, được biểu diễn thông qua ma trận đánh giá *R* (Hình 1.2). Khi đó qui trình xây dựng hệ tư vấn sử dụng phương pháp lọc cộng tác được cụ thể hóa theo Hình 1.5 sau.



Hình 1.5. Tiến trình xử lý của hệ tư vấn sử dụng lọc cộng tác [12]

Tiếp cận phương pháp lọc cộng tác trong xây dựng hệ tư vấn được xem là tiếp cận rất thành công để xây dựng các hệ tư vấn thực tế với một số ưu điểm như đơn giản trong cài đặt và có thể lọc được mọi loại thông tin, đặc biệt đối với thông tin đa phương tiện (ví dụ hình ảnh, âm thanh…) mà không cần phải biểu diễn dưới dạng văn bản.

Một số các nghiên cứu phổ biến đã thực hiện khảo sát, phân loại, cũng như thực nghiệm, đánh giá các thuật toán lọc cộng tác. 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). Mỗi phương pháp tiếp cận có những ưu điểm và hạn chế riêng, khai thác các mối quan hệ trên ma trận đánh giá [1][5][10].

1.5.1.1. Lọc cộng tác dựa vào bộ nhớ

Phương pháp lọc cộng tác theo bộ nhớ được chia thành một số hướng tiếp cận [13][5]: *1) Lọc cộng tác theo bộ nhớ dựa vào người dùng (User-Based Collaborative Filtering)*; *2) Lọc cộng tác theo bộ nhớ dựa vào sản phẩm (Item-Based Collaborative Filtering)*; *3) Đánh trọng số cho mức độ tương tự (Significance Weighting)*; *4) Thiết lập giá trị đánh giá mặc định (Default Voting)*; *5) Chuẩn hóa đánh giá của người dùng trong ma trận đánh giá thông qua ước lượng tần suất xuất hiện ngược của người dùng (Inverse User Frequency)*; *6) Khuếch đại độ tương tự của những người dùng láng giềng với người dùng hiện thời (Case Amplification)*. Mỗi hướng tiếp cận có quá trình xử lý khác nhau đối với dữ liệu đầu vào là ma trận đánh giá để phục vụ cho mục đích tư vấn. Nội dung dưới đây sẽ trình bày chi tiết về những phương pháp này.

❖ **Lọc cộng tác theo bộ nhớ dựa vào người dùng**

Ý tưởng của phương pháp lọc cộng tác theo bộ nhớ dựa vào người dùng [14][13] là sử dụng toàn bộ ma trận đánh giá để chọn ra một tập người dùng tương tự nhất với người dùng hiện thời. Tiếp đó, kết hợp các đánh giá của tập những người dùng tương tự nhất này để đưa ra dự đoán đánh giá cho người dùng hiện thời với sản phẩm chưa biết. Phương pháp được thực hiện theo bốn bước:

- Bước 1. Tính toán mức độ tương tự của tất cả người dùng trong hệ thống với người dùng hiện thời (u_a).
- Bước 2. Xác định tập người dùng láng giềng với u_a bằng việc chọn K_1 người dùng có mức độ tương tự cao nhất với u_a.
- Bước 3. Sinh dự đoán đánh giá của u_a với sản phẩm chưa đánh giá bằng việc kết hợp các đánh giá của các người dùng trong tập láng giềng.

- Bước 4. Tư vấn K_2 sản phẩm mới có mức độ phù hợp cao nhất cho u_a.

Cách tính chi tiết cho bước 1 và bước 3 được miêu tả như sau:

Bước 1. Mức độ tương tự giữa người dùng $u_i \in U$ với người dùng hiện thời u_a, kí hiệu là $sim(u_a, u_i)$, được xem xét dựa vào tập sản phẩm cả hai người dùng đều đánh giá. Có nhiều độ đo khác nhau tính toán mức độ tương tự [1][14] như: Độ đo khoảng cách Euclid, Minkowski…; Độ đo tương tự Cosin, Entropy…; Độ đo tương quan Pearson, Spearman, Kendal,…. Trong đó hai độ đo phổ biến nhất được sử dụng là độ tương quan Pearson và giá trị Cosin giữa hai véc tơ.

- *Độ tương quan Pearson giữa hai người dùng* u_a và u_i :

$$sim_{Pearson}(u_a, u_i) = \frac{\sum_{p_x \in P_{ai}}(r_{ax} - \overline{r_a})(r_{ix} - \overline{r_i})}{\sqrt{\sum_{p_x \in P_{ai}}(r_{ax} - \overline{r_a})^2 \sum_{p_x \in P_{ai}}(r_{ix} - \overline{r_i})^2}} \quad (1.1)$$

Trong đó :

  - $P_{ai} = \{p_x | r_{ax} \neq \emptyset \cap r_{ix} \neq \emptyset\}$ là tập tất cả các sản phẩm cùng được đánh giá bởi u_a và u_i.
  - $\overline{r_a}, \overline{r_i}$ là trung bình cộng các đánh giá khác 0 của u_a và u_i.

- *Độ tương tự Cosin giữa hai người dùng* u_a và u_i: là giá trị Cosin giữa hai véc tơ u_a và u_i theo công thức (1.2). Trong đó, u_a và u_i được biểu diễn bằng véc tơ m chiều ($m = |P_{ai}|$ là số lượng các sản phẩm cả hai người dùng cùng đánh giá).

$$sim_{Cosin}(u_a, u_i) = cos(u_a, u_i) = \frac{\overrightarrow{u_a}.\overrightarrow{u_i}}{||\overrightarrow{u_a}||^2||\overrightarrow{u_i}||^2} = \frac{\sum_{p_x \in P_{ai}} r_{ax}.r_{ix}}{\sqrt{\sum_{p_x \in P_{ai}} r_{ax}^2} \cdot \sqrt{\sum_{p_x \in P_{ai}} r_{ix}^2}} \quad (1.2)$$

Bước 3. Sinh dự đoán đánh giá của u_a với sản phẩm chưa đánh giá p_y :

Gọi $\hat{U}$ là tập K_1 người dùng tương tự nhất đối với u_a (Tập láng giềng với u_a). Khi đó mức độ phù hợp của người dùng u_a với sản phẩm mới p_y được xác định như

một hàm các đánh giá của tập láng giềng theo một số phương pháp phổ dụng dưới đây:

$$\begin{aligned}&\text{(a) } r_{ay} = \frac{1}{K_1}\sum_{u\prime\in\hat{U}} r_{u'y}\\&\text{(b) } r_{ay} = h\sum_{u\prime\in\hat{U}} sim(u_a, u') \times r_{u'y}\\&\text{(c) } r_{ay} = \overline{r_a} + h\sum_{u\prime\in\hat{U}} sim(u_a, u') \times (r_{u'y} - \overline{r_{u'}})\end{aligned} \tag{1.3}$$

Trong đó : h được gọi là nhân tố chuẩn hóa, $\overline{r_a}$ là trung bình các đánh giá của người dùng u_a được xác định theo công thức (1.4).

$$\begin{aligned}h &= 1/\sum_{u\prime\in\hat{U}} sim(u_a, u')\\\overline{r_a} &= \frac{1}{|P_a|}\sum_{p_x\in P_a} r_{ax}\\P_a &= \{p_x \in P \mid r_{ax} \neq 0\}\end{aligned} \tag{1.4}$$

❖ **Lọc cộng tác theo bộ nhớ dựa vào sản phẩm**

Ý tưởng của phương pháp này được thực hiện bằng việc thay vì tính mức độ tương tự giữa các người dùng trong hệ thống với người dùng hiện thời u_a, hệ tư vấn sẽ tính toán mức độ tương tự giữa sản phẩm cần dự đoán đánh giá bởi u_a với các sản phẩm đã được u_a đánh giá. Việc tính toán mức độ tương tự giữa hai sản phẩm được xem xét dựa vào tập người dùng cùng đánh giá cả hai sản phẩm đó. Trên cơ sở đó chọn ra một tập sản phẩm láng giềng với sản phẩm cần dự đoán đánh giá bởi u_a. Kết hợp các đánh giá của u_a với tập sản phẩm láng giềng này để đưa ra dự đoán đánh giá của u_a với sản phẩm cần dự đoán.

Tương tự như việc tính mức độ tương tự giữa hai người dùng cho lọc cộng tác theo bộ nhớ dựa vào người dùng, việc tính toán mức độ tương tự giữa hai sản phẩm cũng được thực hiện tương tự theo một số độ đo đã đề cập. Trong đó có hai độ đo phổ biến nhất được sử dụng là độ tương quan Pearson và giá trị Cosin giữa hai véc tơ.

- *Độ tương quan Pearson giữa hai sản phẩm* p_x và p_y:

$$sim_{Pearson}(p_x, p_y) = \frac{\sum_{u_i \in U_{xy}}(r_{ix} - \overline{r_x})(r_{iy} - \overline{r_y})}{\sqrt{\sum_{u_i \in U_{xy}}(r_{ix} - \overline{r_x})^2 \sum_{u_i \in U_{xy}}(r_{iy} - \overline{r_y})^2}} \quad (1.5)$$

Trong đó :

  - $U_{xy} = \{u_i \mid r_{ix} \neq 0 \cap r_{iy} \neq 0\}$ là tập tất cả các người dùng cùng đánh giá sản phẩm p_x và sản phẩm p_y.
  - $\overline{r_x}, \overline{r_y}$ là trung bình cộng các đánh giá khác 0 cho p_x và p_y.

- *Độ tương tự Cosin giữa sản phẩm* p_x và p_y: là cosin của hai véc tơ p_x và p_y theo công thức (1.6). Trong đó, hai sản phẩm p_x và p_y được xem xét như véc tơ n chiều ($n = |U_{xy}|$ là số lượng các người dùng cùng đánh giá sản phẩm p_x và p_y).

$$sim_{Cosin}(p_x, p_y) = cos(p_x, p_y) = \frac{\overrightarrow{p_x}.\overrightarrow{p_y}}{||\overrightarrow{p_x}||^2||\overrightarrow{p_y}||^2} = \frac{\sum_{u_i \in U_{xy}} r_{ix}.r_{iy}}{\sqrt{\sum_{u_i \in U_{xy}} r_{ix}^2} . \sqrt{\sum_{u_i \in U_{xy}} r_{iy}^2}} \quad (1.6)$$

Việc sinh dự đoán đánh giá của người dùng hiện thời u_a với sản phẩm chưa đánh giá p_y được tính toán bằng việc kết hợp các đánh giá của u_a với các sản phẩm trong tập láng giềng của p_y. Gọi $\hat{P}$ là tập gồm K_1 sản phẩm tương tự nhất đối với p_y. Khi đó mức độ phù hợp của u_a với p_y được xác định như một hàm các đánh giá của tập láng giềng. Phương pháp phổ dụng nhất để dự đoán mức độ phù hợp của sản phẩm p_y đối với người dùng u_a được xác định theo công thức (1.7)

$$r_{ay} = \frac{\sum_{p' \in \hat{P}} r_{ap'}.sim(p', p_y)}{\sum_{p' \in \hat{P}} |sim(p', p_y)|} \quad (1.7)$$

❖ **Đánh trọng số cho mức độ tương tự**

Việc tính mức độ tương tự giữa các người dùng $u_i \in U$ với người dùng hiện thời u_a được xem xét dựa vào tập sản phẩm cả hai người dùng đều đánh giá. Trong một số trường hợp số lượng sản phẩm cả u_i và u_a cùng đánh giá là rất ít nhưng giá

trị tương tự giữa u_i và u_a thu được vẫn khá cao nên u_i vẫn được chọn vào tập láng giềng của u_a, điều này dẫn tới chất lượng dự đoán không cao trong nhiều trường hợp. Thực nghiệm cũng chứng minh nếu số lượng sản phẩm cùng đánh giá bởi hai người dùng u_i và u_a lớn thì mức độ tương tự giữa u_i và u_a thu được ổn định hơn và do vậy tập láng giềng của u_a thu được là hữu ích cho việc dự đoán đánh giá [5].

Từ những nghiên cứu và thực nghiệm như vậy, một số nghiên cứu đã đưa ra hướng giải quyết để hạn chế ảnh hưởng của số lượng ít sản phẩm cả hai người dùng cùng đánh giá lên độ chính xác của hệ tư vấn, bằng việc đưa ra một tham số (Significance Weighting) nhằm đánh trọng số cho các mức độ tương tự tính toán được. Cụ thể, nếu số lượng sản phẩm cả hai người dùng u_i và u_a cùng đánh giá (Kí hiệu là z) ít hơn 1 ngưỡng θ thì mức độ tương tự giữa u_i và u_a tính trước đó sẽ được tính lại bằng cách nhân với $\frac{z}{\theta}$. Trong trường hợp $z \geq \theta$ thì mức độ tương tự giữa u_i và u_a giữ nguyên [5]. Công thức (1.8) thể hiện điều này.

$$sim(u_a, u_i) = \begin{cases} sim(u_a, u_i).\frac{z}{\theta} & If\ z < \theta \\ sim(u_a, u_i) & If\ z \geq \theta \end{cases} \quad (1.8)$$

❖ **Thiết lập giá trị đánh giá mặc định**

Một hướng tiếp cận khác được đưa ra để hạn chế ảnh hưởng của số lượng ít sản phẩm cả hai người dùng cùng đánh giá lên độ chính xác của hệ tư vấn, bằng việc thiết lập giá trị đánh giá mặc định của người dùng với các sản phẩm chưa có đánh giá trước đó. Khi đó mức độ tương tự giữa hai người dùng sẽ được xác định dựa trên tập sản phẩm cả hai người dùng đều đánh giá có số lượng khá lớn. Breese, Heckerman và Kadie [15] chỉ ra rằng độ chính xác của hệ tư vấn dựa vào phương pháp lọc cộng tác cải thiện đáng kể khi các giá trị đánh giá chưa biết được thiết lập giá trị mặc định.

❖ **Ước lượng tần suất xuất hiện ngược của người dùng**

Trong hệ thống, những sản phẩm được đánh giá bởi tất cả người dùng nhiều khi không hữu ích cho việc đưa ra tư vấn bằng những sản phẩm không được đánh

giá bởi tất cả người dùng. Chính vì vậy, tần suất xuất hiện ngược của người dùng $f_i = \log(\frac{n}{n_i})$ được sử dụng cho phép ta chú ý nhiều hơn đến những người dùng không đánh giá tất cả sản phẩm của hệ thống (Với n_i là số lượng người dùng đã đánh giá sản phẩm i, n là tổng số lượng người dùng). Theo đó, các đánh giá của người dùng trong ma trận đánh giá được chuẩn hóa lại bằng tích giá trị đánh giá ban đầu với f_i [5][15].

❖ **Khuếch đại mức độ tương tự của những người dùng láng giềng với người dùng hiện thời**

Nhằm khuếch đại mức độ tương tự của những người dùng láng giềng u_i với u_a, Breese, Heckerman và Kadie [15] đề xuất hằng số khuếch đại ρ ($\rho \geq 1$), nhằm chuyển đổi mức độ tương tự giữa hai người dùng u_a và u_i theo công thức (1.9).

$$sim'(u_a, u_i) = sim(u_a, u_i).|sim(u_a, u_i)|^{\rho-1} \quad (1.9)$$

Mức độ tương tự chuyển đổi này sẽ được dùng cho quá trình dự đoán đánh giá của người dùng hiện thời u_a với các sản phẩm chưa được đánh giá.

1.5.1.2. Lọc cộng tác dựa vào mô hình

Khác với phương pháp lọc cộng tác dựa vào bộ nhớ, phương pháp lọc cộng tác dựa vào mô hình [1][16] sử dụng ma trận đánh giá để xây dựng mô hình dự đoán sinh ra tư vấn cho người dùng. Ưu điểm của phương pháp này là mô hình huấn luyện có kích thước nhỏ hơn rất nhiều so với ma trận đánh giá và thực hiện dự đoán nhanh. Mô hình chỉ cần cập nhật lại khi có những thay đổi lớn và chỉ thực hiện lại pha xây dựng mô hình.

Trong cách tiếp cận này, lọc cộng tác có thể sử dụng các kỹ thuật học máy hoặc khai phá dữ liệu như: luật kết hợp, phân cụm, SVM, cây quyết định, mạng nơ ron nhân tạo và học sâu, đồ thị, phân loại và hồi qui, mạng Bayes, thừa số hóa ma trận, sử dụng Ontology và kỹ thuật giảm chiều dữ liệu... để xây dựng mô hình dự đoán [12][16]. Trong phần dưới đây sẽ trình bày tóm tắt về một số mô hình này.

❖ **Mô hình luật kết hợp**

Với dữ liệu đầu vào cho lọc cộng tác là ma trận đánh giá, mô hình luật kết hợp [17] áp dụng các thuật toán khai phá luật kết hợp nhằm trích xuất ra những luật dự đoán sự xuất hiện những sản phẩm tư vấn dựa trên mối liên hệ của nó với các sản phẩm khác của hệ thống. Luật kết hợp được biểu diễn dưới dạng $A \rightarrow B$, trong đó A, B đại diện cho 2 sản phẩm của hệ thống. Mô hình luật kết hợp được đánh giá là cải thiện hiệu quả hiệu năng và không gian lưu trữ của hệ tư vấn trong một số trường hợp, tuy nhiên khi số lượng người dùng và sản phẩm lớn thì việc tìm ra các luật kết hợp cho hệ tư vấn sẽ trở lên khá phức tạp.

❖ **Mô hình phân cụm**

Kỹ thuật phân cụm [18][19] được thực hiện bằng cách chia các đối tượng dữ liệu ban đầu vào trong các cụm dữ liệu khác nhau. Trong đó, một cụm là tập các đối tượng dữ liệu có các phần tử trong cụm giống nhau nhiều nhất và khác nhau nhiều nhất đối với các phần tử thuộc các cụm khác. Điều này nhằm 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.

Áp dụng các phương pháp phân cụm cho lọc cộng tác để phân chia tập người dùng (hoặc tập sản phẩm) thành các cụm người dùng (hoặc cụm sản phẩm) có sở thích tương tự nhau. Khi đó, người dùng (hoặc sản phẩm) thuộc cụm nào sẽ được dự đoán và tư vấn các sản phẩm được đánh giá cao dựa vào cụm đó [20].

Một hướng tiếp cận phân cụm khác nhóm đồng thời cho cả người dùng và sản phẩm thuộc cùng một cụm, điều này cho phép mỗi người dùng hoặc sản phẩm có thể thuộc nhiều cụm khác nhau, khi đó đánh giá dự đoán của người dùng hiện thời cho các sản phẩm mới sẽ được tính trung bình trên các cụm dữ liệu [21]. K-means và SOM (Self-Organizing Map) là hai phương pháp điển hình cho hướng tiếp cận này. Trong đó, phương pháp K-means tiến hành chia tập N sản phẩm vào K cụm [18], phương pháp học không giám sát SOM có nền tảng dựa trên kỹ thuật phân cụm nơ ron nhân tạo [22].

❖ **Mô hình SVM**

Mô hình máy véc tơ hỗ trợ SVM (Support Vector Machine) [23] là mô hình phân loại dữ liệu nhị phân và có thể mở rộng cho phân loại đa lớp, trong đó mỗi điểm dữ liệu x_i được biểu diễn dưới dạng một véc tơ n-chiều. SVM thực hiện trên nguyên tắc xây dựng một siêu phẳng (hyperplane), dạng $W^T X = C$, đóng vai trò là ranh giới giữa hai nhóm điểm dữ liệu đã biết. Siêu phẳng được lựa chọn sao cho khoảng cách từ các điểm dữ liệu thuộc hai lớp tới nó là xa nhất có thể, điều này giúp sai số phân loại của mô hình SVM càng bé. Trên cơ sở siêu phẳng định nghĩa được, mô hình SVM sẽ xác định lớp cho điểm dữ liệu mới căn cứ vào vị trí của nó ở trên hay dưới siêu phẳng.

Áp dụng mô hình SVM cho lọc cộng tác [23] trên cơ sở xây dựng một siêu phẳng từ tập dữ liệu đánh giá đã biết giúp phân tách những sản phẩm có thể thích và không thích bởi người dùng. Mức độ ưa thích này được dự đoán căn cứ vào vị trí và khoảng cách từ điểm dữ liệu biểu diễn sản phẩm tới siêu phẳng, từ đó đưa ra tư vấn sản phẩm phù hợp với người dùng hiện thời. Hiệu quả của mô hình SVM cho lọc cộng tác được đánh giá là phụ thuộc vào phương pháp biểu diễn các điểm dữ liệu và phương pháp tối ưu tham số trong việc xác định siêu phẳng.

❖ **Mô hình cây quyết định**

Cây quyết định [24] là một cấu trúc ra quyết định có dạng đồ thị hình cây. Với đầu vào là tập dữ liệu huấn luyện đã được gán nhãn biết trước, cây quyết định sẽ phân tích và đưa ra nhãn phân loại cho một ví dụ mới. Áp dụng cây quyết định trong việc dự đoán đánh giá của người dùng hiện thời cho các sản phẩm chưa biết được cho là trực quan và dễ hiểu hơn các phương pháp phân loại khác như Support Vector Machine (SVM) và Neural Networks.

❖ **Mô hình phân loại và hồi qui**

Cho tập hợp gồm N véc tơ M chiều $\{x_i\}$. Mục tiêu của phân loại hay hồi qui là dự đoán chính xác giá trị đầu ra $\{c_i\}$ tương ứng với đầu vào. Trong trường hợp phân loại, c_i nhận một giá trị từ một tập hữu hạn gọi là tập các nhãn. Trong trường hợp

hồi qui, c_i có thể nhận một giá trị thực. Áp dụng mô hình phân loại cho lọc cộng tác [25][26], mỗi sản phẩm (hoặc người dùng) được xây dựng một bộ phân loại riêng. Bộ phân loại cho sản phẩm y phân loại tập người dùng dựa trên những người dùng khác đã đánh giá sản phẩm y. Các bộ phân loại được tiến hành huấn luyện độc lập nhau trên tập các ví dụ huấn luyện.

❖ **Mô hình mạng Bayes**

Mô hình mạng Bayes biểu diễn mỗi sản phẩm như một đỉnh của đồ thị, trạng thái của đỉnh tương ứng với giá trị đánh giá đã biết của người dùng đối với sản phẩm. Cấu trúc của mạng được nhận biết từ tập dữ liệu huấn luyện.

Breese, Heckerman và Kadie [15] đề xuất phương pháp mạng Bayes đơn giản cho lọc cộng tác theo công thức (1.10). Breese giả thiết các giá trị đánh giá được xem xét như những số nguyên nằm giữa 0 và n. Đánh giá chưa biết của người dùng u đối với sản phẩm p là $r_{u,p}$ được ước lượng thông qua những đánh giá trước đó của người dùng u. Gọi $P_u = \{ p' \in P \mid r_{u,p'} \neq 0\}$. Khi đó, đánh giá chưa biết của người dùng u đối với sản phẩm p được tính theo công thức (1.10).

$$r_{u,p} = E(r_{u,p}) = \sum_{i=0}^{n} i \times \Pr(r_{u,p} = i | r_{u,p'}, p' \in P_u) \tag{1.10}$$

Billsus và Pazzani[27] chuyển đổi dữ liệu có nhiều mức đánh giá thành dữ liệu nhị phân. Khi đó, ma trận đánh giá được chuyển đổi thành ma trận bao gồm đặc trưng nhị phân. Việc chuyển đổi này làm cho việc sử dụng mô hình mạng Bayes trở nên thuận tiện hơn. Tuy nhiên, kết quả phân loại theo các đặc trưng nhị phân không phản ánh đúng các bộ dữ liệu thực. Su và Khoshgoftaar [28] mở rộng mô hình mạng Bayes cho các tập dữ liệu thực gồm nhiều lớp đánh giá khác nhau. Kết quả dự đoán của mô hình tốt hơn so với các phương pháp dựa trên độ tương quan Pearson và mô hình mạng Bayes đơn giản.

❖ **Mô hình thừa số hóa ma trận**

Trên thực tế ma trận đánh giá của lọc cộng tác khá thưa do mỗi người dùng chỉ đưa ra một số ít đánh giá của mình cho các sản phẩm của hệ thống [1][10], dẫn tới hệ tư vấn dựa vào phương pháp lọc cộng tác cho lại hiệu quả không cao. Các mô hình nhân tố tiềm ẩn (Latent Factor Model), trong đó mô hình thừa số hóa ma trận (MF - Matrix Factorization) là một điển hình cải thiện đáng kể chất lượng dự đoán của hệ tư vấn [29][10].

Kỹ thuật thừa số hóa ma trận được thực hiện bằng việc chia một ma trận lớn X thành hai ma trận có kích thước nhỏ hơn W và H, sao cho ta có thể xây dựng lại X từ hai ma trận nhỏ hơn này càng chính xác càng tốt [29], nghĩa là $X \sim WH^T$. Trong đó, $W \in R^{|U|\times K}$ là một ma trận mà mỗi dòng là một véc tơ bao gồm K nhân tố tiềm ẩn (latent factors) mô tả người dùng u và $H \in R^{|P|\times K}$ là một ma trận mà mỗi dòng là một véc tơ bao gồm K nhân tố tiềm ẩn mô tả cho sản phẩm p (Lưu ý: $K << |U|$ và $K << |I|$).

Gọi w_{uk} và h_{ik} là các phần tử tương ứng của hai ma trận W và H, khi đó xếp hạng của người dùng u với sản phẩm p được dự đoán bởi công thức (1.11).

$$\hat{r}_{up} = \sum_{k=1}^{K} w_{uk} h_{pk} = w.h^T \tag{1.11}$$

Như vậy vấn đề then chốt của kỹ thuật thừa số hóa ma trận là làm sao để tìm được giá trị của hai tham số W và H. Hai tham số này có được bằng cách tối ưu hóa hàm mục tiêu. Một trong những kỹ thuật có thể dùng để tối ưu hóa hàm mục tiêu là dùng SGD (Stochastic Gradient Descent) [29]. Mô hình thừa số hóa ma trận được đánh giá là cho lại hiệu năng tư vấn cao kể cả trong trường hợp dữ liệu thưa, tuy nhiên khá tốn kém trong việc xây dựng mô hình và cần không gian lưu trữ lớn.

❖ **Mô hình mạng nơ ron nhân tạo và học sâu**

Mạng nơ-ron nhân tạo ANN (Artificial Neural Network) là một mô hình toán học được xây dựng dựa trên cơ sở các mạng nơ-ron sinh học. Nó gồm có một nhóm các nơ-ron nhân tạo (nút) được xếp trong các tầng, các nút được nối từ tầng này tới

tầng khác và các kết nối giữa các nút được đánh trọng số kết nối. Các mạng nơ ron nhân tạo nhiều lớp này còn được gọi là mạng sâu (deep network). Thông tin được xử lý bằng cách truyền theo các kết nối và tính giá trị mới tại các nút nằm ở tầng sau đó. ANN được đánh giá là mô hình học máy mạnh mẽ, nắm bắt được các mối quan hệ phức tạp trong tập hợp dữ liệu và hạn chế nhiễu dữ liệu trong nhiều trường hợp [30]. Tuy nhiên nhược điểm chủ yếu của ANN là nó khó để đưa ra cấu trúc mạng lý tưởng và yêu cầu dữ liệu huấn luyện là tương đối lớn.

Một số nghiên cứu đã ứng dụng mạng nơ ron nhân tạo và các phương pháp học sâu (Deep learning) vào hệ tư vấn để dự đoán đánh giá cho người dùng hiện thời về sản phẩm chưa biết như [30][31][32].

- **Mô hình đồ thị**

Mô hình đồ thị (Graph-based) còn được biết đến là mô hình phân tích liên kết (Link analysis). Đây là mô hình khai thác thông tin dữ liệu thông qua mối liên kết giữa những đối tượng dữ liệu trên đồ thị. Một số giải thuật được áp dụng cho mô hình loại này như PageRank và HITS, trong đó coi mỗi đối tượng dữ liệu như một nút của đồ thị [33]. Ứng dụng mô hình đồ thị này cho hệ tư vấn [34][35] cho phép khai thác các mối quan hệ trực tiếp và gián tiếp giữa các đối tượng dữ liệu (người dùng, sản phẩm) nhằm giải quyết vấn đề dữ liệu thưa của lọc cộng tác, từ đó nâng cao chất lượng tư vấn.

- **Một số mô hình khác** sử dụng Ontology và kỹ thuật giảm chiều dữ liệu SVD (Singular Value Decomposition) [36][37] cho lọc cộng tác giúp giải quyết vấn đề dữ liệu thưa và khả năng mở rộng dữ liệu, từ đó nâng cao độ chính xác và giảm thời gian tính toán.

1.5.1.3. Những vấn đề khi áp dụng phương pháp lọc cộng tác

So với lọc theo nội dung, lọc cộng tác có ưu điểm như đơn giản trong cài đặt và có thể thực hiện tốt trên tất cả các dạng thông tin nhưng gặp phải một số vấn đề sau [1][10] :

- *Vấn đề người dùng mới*: Để phân bổ chính xác các sản phẩm người dùng quan tâm, lọc cộng tác phải ước lượng được sở thích của người dùng đối với các sản phẩm mới thông qua những đánh giá của họ trong quá khứ. Trong trường hợp hệ thống có một người dùng mới, số lượng đánh giá của người dùng này cho các sản phẩm của hệ thống là 0, khi đó phương pháp lọc cộng tác không thể đưa ra những tư vấn chính xác cho người dùng đó.
- *Vấn đề sản phẩm mới*: Trên thực tế các sản phẩm thường xuyên được bổ sung, cập nhật vào hệ thống. Khi xuất hiện một sản phẩm mới, tất cả đánh giá của người dùng cho sản phẩm này đều là 0. Do vậy, lọc cộng tác không thể tư vấn sản phẩm mới này cho bất kỳ người dùng nào trong hệ thống.
- *Vấn đề dữ liệu thưa*: Kết quả dự đoán của lọc cộng tác phụ thuộc chủ yếu vào số lượng đánh giá biết trước của người dùng đối với các sản phẩm. Tuy nhiên, đối với các hệ thống thực tế, số lượng người dùng và sản phẩm là rất lớn (hàng triệu người dùng và sản phẩm) nhưng số lượng những đánh giá biết trước thường rất nhỏ so với số lượng các đánh giá cần dự đoán.
- *Vấn đề sở thích thay đổi theo thời gian*: Hệ thống lọc cộng tác đưa ra tư vấn cho người dùng những sản phẩm của những người có sở thích giống họ đã từng ưa thích trong quá khứ. Khi sở thích của các người dùng đối với các sản phẩm trong hệ thống thay đổi theo thời gian, nếu hệ thống không học lại thì chất lượng tư vấn của hệ thống tới người dùng sẽ giảm chính xác.

1.5.1.4. Ví dụ một số hệ tư vấn sử dụng phương pháp lọc cộng tác

Youtube [38] là hệ thống cho phép người dùng theo dõi, chia sẻ và tư vấn các video trực tuyến cá nhân hóa tới người dùng. Tính năng tư vấn của Youtube được xây dựng sử dụng phương pháp lọc cộng tác dựa vào mô hình để đưa ra tư vấn các video phù hợp với nhu cầu của người dùng. Mô hình được sử dụng ở đây là mạng nơ ron sâu (Deep Neural Networks).

Netflix [39] là hệ thống hệ tư vấn về phim rất lớn sử dụng phương pháp lọc cộng tác dựa vào mô hình thừa số hóa ma trận để đưa ra tư vấn các sản phẩm phù hợp cho người dùng hiện thời.

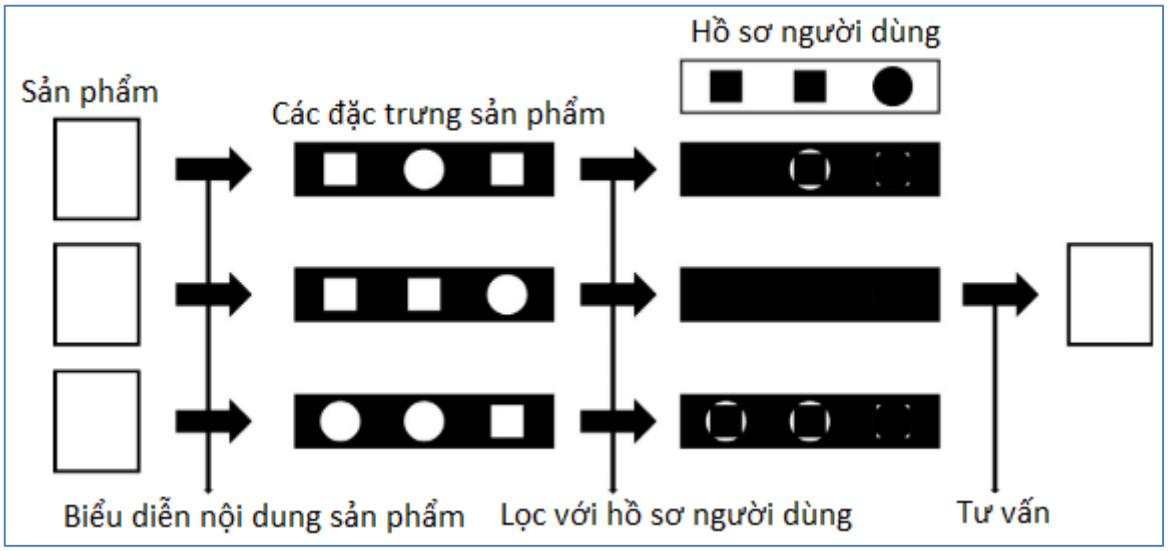
Amazon [21] là hệ tư vấn thương mại điện tử rất nổi tiếng sử dụng phương pháp lọc cộng tác theo bộ nhớ dựa vào sản phẩm. Hệ tư vấn này khai thác thông tin đầu và là ma trận đánh giá thu được thông qua các đánh giá tường minh của người dùng với sản phẩm để huấn luyện và đưa ra tư vấn.

1.5.2. Hệ tư vấn sử dụng lọc theo nội dung

Ý tưởng của hệ tư vấn sử dụng phương pháp lọc theo nội dung là 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 [1][2][10]. Các phương pháp dựa trên tiếp cận nội dung thông thường sẽ thực hiện các bước sau:

- **Bước 1.** Biểu diễn nội dung của đối tượng khuyến nghị $p \in P$, được ký hiệu là $Content(p)$, thông qua tập |C| đặc trưng nội dung của p. Tập các đặc trưng của sản phẩm p được xây dựng bằng các kỹ thuật truy vấn thông tin.
- **Bước 2.** Mô hình hóa sở thích người dùng $u \in U$, gọi tắt là hồ sơ người dùng, ký hiệu $UserProfile(u)$. Hồ sơ của người dùng u thực chất là lịch sử truy cập hoặc đánh giá của người đó đối với các đặc trưng nội dung sản phẩm. $UserProfile(u)$ được xây dựng bằng cách phân tích nội dung các sản phẩm mà người dùng u đã từng truy nhập hoặc đánh giá dựa trên các kỹ thuật truy vấn thông tin. Do vậy $UserProfile(u)$ là một véc tơ biểu diễn thông qua $|C|$ đặc trưng nội dung của P.
- **Bước 3.** Dự đoán đánh giá của người dùng u với sản phẩm p dựa trên độ tương tự nội dung của p với hồ sơ người dùng u. Hệ thống sẽ ưu tiên tư vấn những đối tượng p có nội dung tương tự cao nhất với hồ sơ người dùng u.

Tiến trình xử lý của hệ tư vấn sử dụng phương pháp lọc theo nội dung được cụ thể hóa trong Hình 1.6 sau.



Hình 1.6. Tiến trình xử lý của hệ tư vấn sử dụng lọc theo nội dung [2]

Các phương pháp tư vấn truyền thống dựa trên nội dung có thể chia thành hai nhóm chính: 1) Lọc theo nội dung dựa vào bộ nhớ, thực hiện tính toán độ tương tự giữa $Content(p)$ và $UserProfile(u)$ dùng các độ đo tương tự như Cosine, Euclide...; 2) Lọc theo nội dung dựa vào mô hình, với mô hình được học từ dữ liệu dùng các kỹ thuật thống kê hoặc học máy để phân các đối tượng khuyến nghị thành những đối tượng người dùng quan tâm hay không quan tâm. Nội dung cụ thể được trình bày sau đây.

1.5.2.1. Lọc theo nội dung dựa vào bộ nhớ

Lọc theo nội dung dựa vào bộ nhớ là phương pháp sử dụng toàn bộ tập hồ sơ sản phẩm hoặc tập hồ sơ người dùng để thực hiện huấn luyện và dự đoán.

Nội dung phần này sẽ trình bày về phương pháp lọc theo nội dung dựa vào bộ nhớ sử dụng toàn bộ hồ sơ người dùng, những sản phẩm mới có mức độ tương tự cao nhất với hồ sơ người dùng sẽ được dùng để tư vấn cho người dùng này. Hệ tư vấn xây dựng theo hướng này được biết đến là hệ tư vấn theo nội dung sản phẩm.

Với phương pháp lọc theo nội dung dựa vào bộ nhớ sử dụng toàn bộ hồ sơ sản phẩm được thực hiện tương tự. Hệ tư vấn xây dựng theo hướng này được biết đến là hệ tư vấn theo nội dung người dùng.

Hệ tư vấn theo nội dung sản phẩm được hiện thực hóa qua ba bước theo Hình 1.6 như sau:

- ***Bước 1. Biểu diễn sản phẩm mới dưới dạng véc tơ trọng số các đặc trưng nội dung sản phẩm***

Phương pháp ước lượng trọng số các đặc trưng nội dung thông dụng nhất thường được sử dụng là phép đo tần suất kết hợp với tần suất xuất hiện ngược (TF/IDF). Phương pháp được thực hiện như sau:

Gọi $f_{i,j}$ là số lần đặc trưng nội dung c_i xuất hiện trong sản phẩm p_j. Khi đó tần suất $TF_{i,j}$ của đặc trưng nội dung c_i trong sản phẩm p_j được xác định theo công thức (1.12).

$$TF_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} \tag{1.12}$$

Ở đây, $max_z f_{z,j}$ là số lần xuất hiện nhiều nhất của đặc trưng nội dung c_z trong sản phẩm p_j.

Tuy nhiên, những đặc trưng nội dung xuất hiện trong nhiều sản phẩm không được dùng để xem xét mức độ tương tự giữa các sản phẩm, thậm chí những đặc trưng nội dung này không chứa đựng nhiều thông tin phản ánh nội dung sản phẩm. Chính vì vậy, tần suất xuất hiện ngược IDF_i kết hợp với tần suất $TF_{i,j}$ cho phép ta chú ý nhiều hơn đến những đặc trưng nội dung có trong sản phẩm này nhưng ít xuất hiện trong các sản phẩm khác.

Phương pháp xác định tần suất xuất hiện ngược được thực hiện như sau: Giả sử hệ thống có N sản phẩm cần được phân bổ hoặc tư vấn cho người dùng và đặc trưng nội dung c_i xuất hiện trong n_i sản phẩm. Tần suất xuất hiện ngược IDF_i của đặc trưng nội dung c_i có tần suất xuất hiện trong sản phẩm p_j là $TF_{i,j}$ được xác định theo công thức (1.13), trọng số các đặc trưng nội dung c_i được xác định theo công thức (1.14).

$$IDF_i = log\frac{N}{n_i} \tag{1.13}$$

$$w_{i,j} = TF_{i,j} \times IDF_i \tag{1.14}$$

Trong công thức (1.14), nếu thay $n_i \cong N$ hay đặc trưng nội dung c_i xuất hiện trong đại đa số các sản phẩm thì trọng số $w_{i,j} \cong 0$, có nghĩa là những đặc trưng nội dung có trong mọi sản phẩm thì đặc trưng đó không chứa nhiều nội dung thông tin phản ánh sản phẩm. Ngược lại, nếu đặc trưng nội dung chỉ xuất hiện trong một sản phẩm thì $n_i = 1$, khi đó $w_{i,j} = TF_{i,j}$. Do đó, những đặc trưng nội dung chỉ xuất hiện ở một loại sản phẩm và không xuất hiện ở những sản phẩm khác thì những đặc trưng nội dung này chứa nhiều nội dung quan trọng đối với sản phẩm.

Bằng cách ước lượng này, mỗi sản phẩm $p_x \in P$ được biểu diễn như một véc tơ trọng số các đặc trưng nội dung $w_x = \{w_{1x}, w_{2x}, .., w_{|C|x}\}$. Trong đó $|C|$ là số lượng đặc trưng nội dung của toàn bộ sản phẩm.

- ***Bước 2. Biểu diễn hồ sơ người dùng dưới dạng véc tơ trọng số các đặc trưng nội dung sản phẩm***

Với mỗi người dùng $u_i \in U$, $w_i = \{w_{i1}, w_{i2}, \dots, w_{i|c|}\}$ là véc tơ trọng số các đặc trưng nội dung sản phẩm $c_s \in C$ đối với mỗi người dùng u_i, hay còn được gọi là hồ sơ của người dùng u_i. Trong đó mỗi w_{is} là trọng số, biểu diễn mức độ quan trọng của đặc trưng nội dung c_s đối với người dùng u_i. Véc tơ w_i được tính toán bằng các kỹ thuật khác nhau từ véc tơ trọng số các đặc trưng nội dung của sản phẩm đã được người dùng thường xuyên truy cập hoặc đánh giá. Balanovic [40] tính toán véc tơ trọng số mỗi hồ sơ người dùng w_i bằng cách lấy trung bình cộng véc tơ trọng số w_x trên các sản phẩm $p_x \in P$ mà người dùng đã từng truy cập hoặc đánh giá. Pazzani [41] sử dụng bộ phân loại Bayes ước lượng khả năng giống nhau của sản phẩm và đề xuất thuật toán Winnow thực hiện trong những trường hợp có nhiều đặc trưng nội dung.

- ***Bước 3. Tính độ tương tự giữa sản phẩm mới và hồ sơ người dùng***

Theo cách biểu diễn như trên, sản phẩm mới và hồ sơ người dùng đều được biểu diễn dưới dạng véc tơ trọng số các đặc trưng nội dung sản phẩm có cùng chiều và ước lượng theo cùng một phương pháp (trong trường hợp này là TF-IDF). Do vậy, mức độ phù hợp của sản phẩm $p_x \in P$ với người dùng $u_i \in U$ được xác định căn cứ vào mức độ tương tự giữa sản phẩm mới p_x và hồ sơ của người dùng u_i. Những sản phẩm có mức độ tương tự cao nhất với hồ sơ người dùng hiện thời sẽ được dùng để tư vấn cho người dùng hiện thời.

Phương pháp phổ biến để ước lượng mức độ tương tự giữa sản phẩm $p_x \in P$ và hồ sơ người dùng $u_i \in U$ là dùng độ đo Cosin giữa hai véc tơ trọng số w_i và w_x.

$$sim_{Cosin}(w_i, w_x) = \frac{\overrightarrow{w_i}.\overrightarrow{w_x}}{||\overrightarrow{w_i}||^2||\overrightarrow{w_x}||^2} = \frac{\sum_{s=1}^{|C|} w_{s,i}.w_{s,x}}{\sqrt{\sum_{s=1}^{|C|} w_{s,i}^2}.\sqrt{\sum_{s=1}^{|C|} w_{s,x}^2}} \quad (1.15)$$

Ở đây, $|C|$ là số lượng đặc trưng nội dung sản phẩm. Trong công thức 1.15, nếu Cosin của hai véc tơ gần tới 1, hay góc tạo bởi hai véc tơ này nhỏ thì mức độ tương tự giữa sản phẩm và hồ sơ người dùng càng cao, khi đó sản phẩm có mức độ phù hợp càng cao với người dùng. Ngược lại, nếu Cosin của hai véc tơ gần bằng 0, hay góc tạo bởi hai véc tơ lớn thì mức độ tương tự giữa sản phẩm và hồ sơ người dùng càng thấp, khi đó sản phẩm có mức độ phù hợp càng thấp với người dùng. Theo cách đo này, nếu người dùng u_a truy cập nhiều sản phẩm liên quan đến một chủ đề nào đó thì hệ thống lọc theo nội dung sẽ phân bổ những sản phẩm của chủ đề đó cho người dùng u_a.

1.5.2.2. Lọc theo nội dung dựa vào mô hình

Lọc theo nội dung dựa vào mô hình là phương pháp sử dụng toàn bộ tập hồ sơ sản phẩm hoặc tập hồ sơ người dùng để thực hiện huấn luyện. Kết quả của mô hình huấn luyện sẽ sử dụng trong mô hình dự đoán để sinh ra tư vấn cho người dùng. Trong cách tiếp cận này, lọc theo nội dung có thể sử dụng các kỹ thuật thống kê, học máy như : mạng Bayes, phân cụm, cây quyết định, mạng nơ ron nhân tạo…[24] để sinh dự đoán cho người dùng.

Pazzani và Billsus [42] sử dụng bộ phân loại Bayes dựa trên những đánh giá '*thích'* hoặc *'không thích'* của người dùng để phân loại các sản phẩm.

Solombo [43] đề xuất mô hình lọc thích nghi, trong đó chú trọng đến việc quan sát mức độ phù hợp của tất cả các sản phẩm.

Zhang [44] đề xuất mô hình tối ưu tập các sản phẩm tương tự dựa vào giá trị ngưỡng. Trong đó, giá trị ngưỡng được ước lượng dựa trên tập sản phẩm thích hợp và tập sản phẩm không thích hợp với mỗi hồ sơ người dùng.

1.5.2.3. Những vấn đề khi áp dụng phương pháp lọc theo nội dung

Mặc dù lọc theo nội dung đã áp dụng thành công cho nhiều ứng dụng lọc văn bản, tuy vậy phương pháp vẫn tồn tại một số vấn đề cần tiếp tục nghiên cứu giải quyết, đó là vấn đề trích chọn đặc trưng và người dùng mới [1][5].

- *Vấn đề trích chọn đặc trưng:* Lọc theo nội dung kế thừa và phát triển dựa chủ yếu vào các phương pháp trích chọn đặc trưng trong lĩnh vực truy vấn thông tin. Để có một tập các đặc trưng đầy đủ, nội dung tài liệu phải được biểu diễn dưới dạng phù hợp để máy tính có thể tự động phân tích, tính toán trọng số các đặc trưng nội dung hoặc phải được thực hiện bán tự động. Phương pháp sẽ khó áp dụng trong những trường hợp việc trích chọn nội dung phức tạp, chẳng hạn trích chọn đặc trưng nội dung các đối tượng dữ liệu đa phương tiện (hình ảnh, âm thanh,...).
- *Vấn đề người dùng mới:* Các hệ thống lọc theo nội dung chỉ thực hiện hiệu quả khi người dùng đánh giá hoặc truy nhập một số lượng sản phẩm đủ lớn. Trong trường hợp người dùng mới, hồ sơ người dùng được biểu diễn dưới dạng véc tơ trọng số các đặc trưng nội dung sản phẩm có các thành phần là 0, vì vậy hệ thống sẽ không thể thực hiện dự đoán và phân bổ những sản phẩm thích hợp cho người dùng.

1.5.2.4. Ví dụ một số hệ tư vấn sử dụng phương pháp lọc theo nội dung

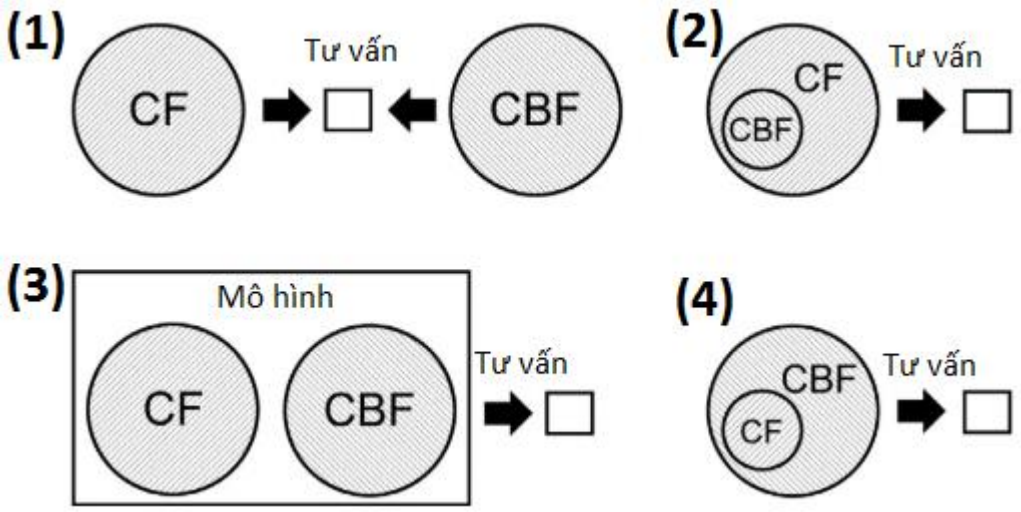
News Dude [27] là một hệ thống tổng hợp giọng nói nhằm đọc truyện cho người dùng. Hệ thống này sử dụng phương pháp TF-IDF để biểu diễn nội dung truyện, sau đó sử dụng độ đo tương tự Cosin để tính mức độ tương tự giữa truyện và hồ sơ người dùng.

LIBRA [45] là hệ tư vấn sách được xây dựng dựa vào phương pháp lọc theo nội dung. Hệ thống này sử dụng mô hình phân loại Naı¨ve Bayes học hồ sơ người dùng là cơ sở để đưa ra tư vấn các sản phẩm phù hợp với người dùng hiện thời. Ưu điểm của hệ tư vấn LIBRA là các tư vấn đưa ra bởi hệ thống là có sơ sở giải thích rõ ràng.

1.5.3. Hệ tư vấn sử dụng lọc kết hợp

Lọc theo nội dung khai thác những khía cạnh liên quan đến các đặc trưng nội dung thông tin của những đối tượng cần lọc. Trái lại, lọc cộng tác khai thác những khía cạnh liên quan đến thói quen sử dụng các loại thông tin khác nhau của mỗi người dùng. Mỗi phương pháp đều có những thế mạnh và hạn chế nhất định, do vậy để phát huy điểm mạnh và hạn chế những điểm yếu của từng kỹ thuật tư vấn riêng lẻ, các phương pháp lọc kết hợp được đưa ra nhằm cải thiện hiệu quả tư vấn sản phẩm mới phù hợp tới người dùng hiện thời [1][12]. Nhiều kết quả so sánh đã chứng tỏ phương pháp lọc kết hợp cho lại kết quả dự đoán tốt hơn so với các phương pháp lọc cộng tác và lọc nội dung thuần túy. Đặc biệt, lọc kết hợp hạn chế ảnh hưởng của vấn đề dữ liệu thưa và người dùng mới [46].

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 [4][47].



Hình 1.7. Các phương pháp kết hợp lọc cộng tác (CF) và lọc nội dung (CBF) [2]

Các phương pháp lọc kết hợp sẽ được trình bày cụ thể dưới đây.

1.5.3.1. Kết hợp các kết quả dự đoán của CF và CBF trong lọc kết hợp [12]

Kết hợp tuyến tính có trọng số

Là phương pháp xây dựng hai mô hình lọc nội dung và lọc cộng tác độc lập nhau. Kết quả đánh giá dự đoán của toàn bộ mô hình được kết hợp với nhau theo một hàm tuyến tính. Ưu điểm của phương pháp này là kế thừa được phương pháp biểu diễn và tính toán vốn có của các phương pháp lọc cơ sở. Nhược điểm lớn nhất của mô hình này là cho lại kết quả không cao vì chưa có sự kết hợp hiệu quả giữa nội dung và đánh giá người dùng. P-tango [48] là hệ tư vấn được xây dựng dựa trên kỹ thuật này.

Kết hợp chuyển đổi

Phương pháp kết hợp chuyển đổi cho phép linh hoạt trong việc lựa chọn phương pháp tư vấn theo lọc cộng tác hay lọc theo nội dung tùy thuộc vào thế mạnh của mỗi phương pháp áp dụng tại từng thời điểm khác nhau của hệ thống.

Ví dụ khi sản phẩm mới tham gia vào hệ thống, nếu áp dụng phương pháp lọc cộng tác thì không thể tư vấn các sản phẩm phù hợp cho người dùng hiện thời, do sản phẩm này chưa có đánh giá bởi bất kỳ người dùng nào trong hệ thống. Ở tình huống này thì hệ thống sẽ chuyển đổi sang phương pháp lọc theo nội dung nhằm

khai thác thế mạnh của phương pháp này. Trong trường hợp khác khi hệ thống thu thập được nhiều đánh giá cho ma trận đánh giá người dùng – sản phẩm thì hệ thống sẽ chuyển đổi sang phương pháp lọc cộng tác nhằm khai thác thế mạnh của lọc cộng tác. Việc chuyển đổi phương pháp là linh động trong từng trường hợp dữ liệu tại các thời điểm khác nhau.

Mặc dù phương pháp kết hợp chuyển đổi tỏ ra khá linh hoạt, phát huy thế mạnh và hạn chế nhược điểm của mỗi phương pháp lọc cơ sở. Tuy nhiên nhược điểm chính của phương pháp này là khá phức tạp trong việc xác định điều kiện và giá trị các tham số cần thiết của hệ thống [47]. DailyLearner [49] là hệ tư vấn sử dụng phương pháp kết hợp chuyển đổi giữa lọc cộng tác và lọc nội dung.

Kết hợp ghép tầng

Phương pháp kết hợp ghép tầng cho phép ghép nối danh sách các sản phẩm tư vấn được đưa ra bởi hai phương pháp tư vấn cơ sở (CF, CBF) trên nguyên tắc danh sách các sản phẩm được đưa ra bởi phương pháp tư vấn đầu tiên sẽ được lọc lại một lần nữa bởi phương pháp tư vấn thứ 2. Phương pháp kết hợp này được đánh giá là khá hiệu quả, đồng thời cho phép kế thừa được phương pháp biểu diễn và tính toán vốn có của các phương pháp tư vấn cơ sở. EntreeC [47] là một ví dụ cho phương pháp này.

Kết hợp hỗn hợp

Phương pháp kết hợp hỗn hợp cho phép thực hiện đồng thời các phương pháp lọc cơ sở, kết quả tư vấn cuối cùng sẽ là tổng hợp kết quả tư vấn của các phương pháp lọc riêng lẻ. Ví dụ của hệ tư vấn sử dụng phương pháp này là hệ thống PTV [50], đây là hệ thống cho phép tư vấn lịch phát sóng truyền hình. Ngoài ra một số hệ tư vấn khác được biết đến cũng sử dụng phương pháp kết hợp hỗn hợp như Profinder [51] và PickAFlick [52].

1.5.3.2. Kết hợp đặc tính của lọc nội dung vào lọc cộng tác [46][53]

Là phương pháp dựa trên các kỹ thuật lọc cộng tác thuần túy nhưng vẫn duy trì hồ sơ người dùng như một tham biến tham khảo khi tính toán sự tương tự giữa các cặp người dùng. Trong trường hợp dữ liệu thưa hoặc người dùng mới, mức độ tương tự giữa hồ sơ người dùng và sản phẩm sẽ được xem xét đến để tạo nên dự đoán.

1.5.3.3. Kết hợp đặc tính của lọc cộng tác vào lọc nội dung [26][54]

Là phương pháp xem xét các đánh giá người dùng của lọc cộng tác như một thành phần trong mỗi hồ sơ người dùng. Phương pháp dự đoán thực hiện theo lọc nội dung thuần túy dựa trên biểu diễn hồ sơ người dùng mở rộng. Phương pháp phổ biến nhất thực hiện theo mô hình này là sử dụng các kỹ thuật giảm số chiều cho hồ sơ người dùng trước khi kết hợp với đánh giá người dùng. Pipper[25] là hệ thống tư vấn phim cho phép kết hợp các đánh giá của lọc cộng tác vào là đặc trưng của lọc nội dung.

1.5.3.4. Kết hợp tăng cường đặc trưng [12]

Phương pháp kết hợp tăng cường là một dạng mạnh hơn của phương pháp kết hợp đặc tính của lọc cộng tác vào lọc nội dung hoặc kết hợp đặc tính của lọc nội dung vào lọc cộng tác. Phương pháp kết hợp này dựa trên dữ liệu đầu vào là tập hợp các đánh giá của người dùng với sản phẩm và một số thông tin khác thu được từ phương pháp tư vấn trước đó, như vậy mô hình học dữ liệu của hệ tư vấn cũng có sự bổ sung điều chỉnh. Ví dụ hệ thống tư vấn sách Libra [45] áp dụng phương pháp tư vấn dựa vào phương pháp lọc theo nội dung với dữ liệu có được từ hệ thống của Amazon, trong đó mô hình học dữ liệu được sử dụng là Nai¨ve Bayes.

1.5.3.5. Xây dựng mô hình hợp nhất giữa lọc cộng tác và lọc nội dung [55][56]

Là phương pháp biểu diễn đặc trưng nội dung và đánh giá người dùng trên cùng một mô hình. Kết quả dự đoán dựa trên mô hình dữ liệu hợp nhất của cả nội dung và đánh giá người dùng. Basu và các cộng sự [25] đề xuất sử dụng lọc cộng

tác và lọc nội dung trong một bộ phân loại đơn lẻ. Schein[57] đề xuất phương pháp thống kê kết hợp hai phương pháp dựa trên mô hình phân tích ngữ nghĩa ẩn *(LSM)*. Ansari[58] đề xuất mô hình hồi qui dựa trên mạng Bayes, trong đó mỗi hồ sơ người dùng và sản phẩm được biểu diễn trong cùng một mô hình thống kê.

1.5.3.6. Những vấn đề còn tồn tại với phương pháp lọc kết hợp

Mặc dù lọc kết hợp được đánh giá là cho phép phát huy ưu điểm và khắc phục nhược điểm của từng kỹ thuật tư vấn riêng lẻ, tuy vậy phương pháp vẫn tồn tại một số vấn đề cần tiếp tục nghiên cứu giải quyết [1][12], đó là :

- *Thiếu sự kết hợp hiệu quả các đặc trưng nội dung vào lọc cộng tác :* Không phải tất cả các đặc trưng nội dung của sản phẩm đều ảnh hưởng đến thói quen sử dụng sản phẩm của tất cả người dùng. Việc tìm ra tập các đặc trưng nội dung có ảnh hưởng quan trọng đến thói quen sử dụng sản phẩm của mỗi người dùng cụ thể sẽ cải thiện đáng kể kết quả dự đoán của các mô hình.
- *Thiếu sự kết hợp hiệu quả các đặc trưng của lọc cộng tác vào lọc nội dung :* Các phương pháp lọc cộng tác thực hiện dự đoán dựa trên tập đánh giá người dùng đối với sản phẩm. Trái lại, các phương pháp lọc nội dung dựa trên biểu diễn nội dung sản phẩm và hồ sơ người dùng. Việc thực hiện tính toán mức độ tương tự theo nội dung trên cả nội dung sản phẩm và đánh giá người dùng chưa giải quyết triệt để mâu thuẫn giữa các cách tiếp cận.
- Cần nâng cao hiệu quả phương pháp biểu diễn và dự đoán cho mô hình hợp nhất cho cả lọc cộng tác và lọc nội dung.

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

Ba phương pháp xây dựng hệ tư vấn theo cách tiếp cận truyền thống trình bày trong Mục 1.5.1, Mục 1.5.2 và Mục 1.5.3 nêu trên tập trung vào khai thác ba loại thông tin đầu vào, gồm người dùng, sản phẩm và phản hồi của người dùng về sản phẩm, để đưa ra gợi ý các sản phẩm mới phù hợp với người dùng hiện thời. Do vậy chất lượng của hệ tư vấn sẽ phụ thuộc vào hiệu quả của phương pháp lọc thông tin

cơ sở (lọc cộng tác, lọc theo nội dung, lọc kết hợp) trên ba loại thông tin đó. Theo phần trình bày ở trên, chúng ta đều thấy mỗi phương pháp theo hướng tiếp cận truyền thống đều có những ưu điểm và hạn chế nhất định cần tiếp tục nghiên cứu giải quyết. Chính vì vậy, 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 [1][4][9].

Theo xu hướng thứ nhất, các phương pháp lọc kết hợp (Hybrid Filtering Recommendation) và các phương pháp lọc dựa trên mối quan tâm (Attention-based Recommendation) đang đặc biệt được quan tâm nghiên cứu.

Cụ thể, một số nghiên cứu về hệ tư vấn theo phương pháp lọc kết hợp được đề cập đến như: Hợp nhất lọc cộng tác và lọc nội dung để giải quyết bài toán tư vấn địa điểm cho các thiết bị di động [59], áp dụng mô hình học sâu cho lọc kết hợp để giải quyết bài toán tư vấn dịch vụ web [60], kết hợp các phương pháp lọc cộng tác khác nhau dựa vào giải thuật phân loại đa lớp để đưa ra tư vấn [61] hay khai thác những tác nhân tiềm ẩn của người dùng và sản phẩm cho hệ tư vấn cộng tác kết hợp dựa trên mô hình mạng nơ ron nhân tạo [62]…Các phương pháp lọc kết hợp mới này được đánh giá là cho độ chính xác tốt, tuy nhiên hầu hết đang tập trung vào một số bài toán tư vấn cụ thể hoặc có độ phức tạp tính toán tương đối lớn.

Đề cập tới các phương pháp lọc dựa trên mối quan tâm, đây là hướng nghiên cứu phát triển các phương pháp lọc tin truyền thống cho hệ tư vấn sử dụng các mô hình học sâu dựa trên mạng nơ ron nhân tạo. Hướng tiếp cận này tập trung vào khai thác tự động những thông tin liên quan nhất tới người dùng hoặc sản phẩm cần tư vấn (vùng quan tâm), thay vì khai thác toàn bộ thông tin của các đối tượng trong hệ thống, từ đó nâng cao hiệu quả tư vấn. Một số nghiên cứu theo hướng tiếp cận này có thể kể đến như: Phương pháp lọc cộng tác dựa trên mối quan tâm tập trung khai thác những phản hồi tường minh của người dùng với sản phẩm [9], lọc cộng tác dựa trên mối quan tâm sử dụng mô hình mạng nơ ron nhân tạo cho phép khai thác

những phản hồi không tường minh khi người dùng tương tác với sản phẩm [63], mô hình hóa đa dạng sở thích của người dùng (Dựa vào toàn bộ lịch sử sử dụng sản phẩm và những vùng quan tâm của người dùng) trong một mạng nơ ron nhân tạo sâu cho hệ tư vấn [64], mô hình tư vấn dựa trên mối quan tâm cụ thể của người dùng với từng đặc trưng sản phẩm (thông qua mạng lưới vùng quan tâm) để giải quyết bài toán dự đoán đánh giá [65], … Các nghiên cứu này chỉ ra lọc dựa trên mối quan tâm nhìn chung được đánh giá là cho độ chính xác tương đối tốt và có cơ sở lý giải được lý do đằng sau các tư vấn, tuy nhiên việc áp dụng các mô hình học sâu yêu cầu lượng dữ liệu huấn luyện là tương đối lớn và độ phức tạp cũng lớn hơn các phương pháp tư vấn truyền thống.

Theo xu hướng thứ hai, ngoài các thông tin cơ bản là người dùng, sản phẩm và phản hồi của người dùng về sản phẩm được khai thác trong hệ tư vấn, cộng đồng nghiên cứu cũng xem xét tích hợp thêm các nguồn thông tin hữu ích khác thu nhận được của hệ thống vào quá trình tư vấn.

Ngày nay các nguồn thông tin thu thập được trong hệ thống tư vấn rất đa dạng, một số thông tin thu được nằm ngoài phạm vi ba loại thông tin nêu trên, ví dụ dữ liệu về mối quan hệ kết bạn (friends), mức độ tin cậy của mối quan hệ từ mạng xã hội (trust),... hoặc thông tin ngữ cảnh sử dụng sản phẩm của người dùng hay nhóm người dùng tham gia. Các thông tin này chủ yếu được thu thập trong các quá trình: 1) Người dùng tương tác với hệ thống; 2) Các hệ thống tương tác với nhau; 3) Thu thập thông qua các thiết bị thu nhận, cảm biến kết nối với hệ thống. Chính vì vậy mà nhu cầu về tính cá nhân hóa với dữ liệu sẽ cao hơn bao giờ hết. Đứng trước thách thức về tính cá nhân hóa dữ liệu tới người dùng lớn như vậy dẫn tới nảy sinh nhu cầu cấp thiết về việc thiết kế, cải tiến và mở rộng các hệ tư vấn truyền thống nhằm khai thác tối đa các thông tin khác liên quan giữa người dùng với sản phẩm, từ đó đưa ra tư vấn các sản phẩm, dịch vụ phù hợp nhất với người dùng hiện thời.

Trong rất đa dạng các dữ liệu liên kết thu thập được bởi hệ tư vấn như vậy, có một loại thông tin bổ sung vào hệ tư vấn được đặc biệt quan tâm nghiên cứu trong

những năm gần đây, đó là thông tin ngữ cảnh sử dụng sản phẩm của người dùng. Chẳng hạn đối với hệ tư vấn du lịch, yếu tố ngữ cảnh có thể là thời gian (buổi trong ngày, thời gian trong tuần, mùa), bạn đồng hành (một mình, gia đình, bạn bè). Những yếu tố này hoàn toàn có thể ảnh hưởng tới quyết định chọn địa điểm du lịch của người dùng. Hệ tư vấn sẽ đóng vai trò ghi nhớ lại sở thích của người dùng theo ngữ cảnh để đưa ra những gợi ý chính xác nhất. Các thông tin ngữ cảnh này có thể đến từ nhiều nguồn, như do người dùng đưa vào hoặc thu thập thông qua các thiết bị, cảm biến kết nối vào Internet, đều là những nguồn thông tin liên quan tới thói quen sử dụng sản phẩm của người dùng. Tất cả những thông tin ngữ cảnh đa dạng này cùng với những thông tin thu thập vốn có trong các hệ tư vấn truyền thống, nếu được khai thác hợp lý sẽ là công cụ hữu ích cung cấp những tư vấn cá nhân hóa mạnh mẽ tới người dùng.

Mặc dù đã có một số đề xuất được đưa ra để giải quyết bài toán tư vấn dựa vào ngữ cảnh, nhưng một số vấn đề điển hình như: vấn đề dữ liệu thưa, độ phức tạp tính toán lớn và làm thế nào để tích hợp hiệu quả thông tin ngữ cảnh vào hệ tư vấn vẫn là vấn đề nghiên cứu mở, có tính thời sự và thu hút được nhiều quan tâm của cộng đồng nghiên cứu. Các kết quả nghiên cứu cũng chỉ ra rằng việc lựa chọn mở rộng các phương pháp tư vấn truyền thống phù hợp sẽ ảnh hưởng đáng kể tới chất lượng của hệ tư vấn dựa vào ngữ cảnh [6].

Trên cơ sở phân tích những vấn đề còn tồn tại của hai xu hướng chính trong nghiên cứu về hệ tư vấn hiện nay là: (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, tác giả tập trung chính vào nghiên cứu đưa ra hai đề xuất giải quyết một số vấn đề điển hình còn tồn tại theo hai xu hướng này, các phân tích và kết quả được trình bày cụ thể trong chương 2 và chương 3. Trong đó, nội dung chương 2, tiếp cận theo xu hướng (2), đưa ra cơ sở lý luận chi tiết, những nghiên cứu liên quan và đề xuất phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh. Động cơ nghiên cứu cho chương 2 nhằm giải quyết vấn đề dữ liệu thưa và tích hợp hiệu quả thông tin ngữ cảnh vào hệ tư vấn.

Nội dung chương 3, tiếp cận theo xu hướng (1), trình bày sâu chuỗi cơ sở lý thuyết để đưa ra đề xuất về phương pháp lọc kết hợp cải tiến từ các phương pháp lọc tin truyền thống nhằm giải quyết vấn đề dữ liệu thưa, đơn giản trong cài đặt và nâng cao độ chính xác tư vấn.

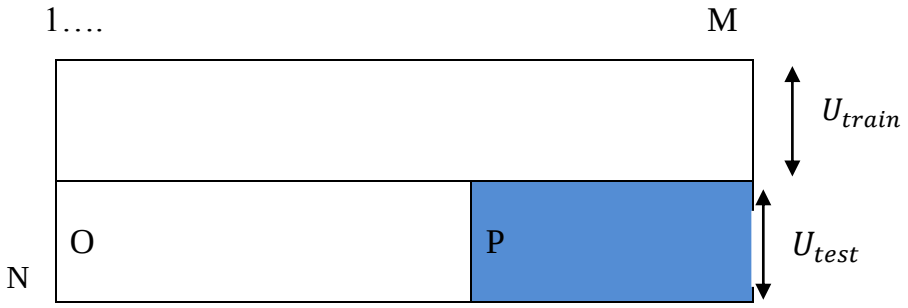
1.6. Các phương pháp và độ đo đánh giá hệ tư vấn

1.6.1. Phương pháp đánh giá hệ thống tư vấn

Trong thực tế, ta cần chọn một mô hình thống kê hoặc học máy phù hợp nhất cho hệ tư vấn của mình. Vấn đề đặt ra là làm thế nào để đánh giá độ chính xác và chọn ra được mô hình phù hợp trong số rất nhiều mô hình đang có hiện nay.

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. Dưới đây là một số cách tiếp cận để chia tập người dùng U thành 2 phần U_{train} *và* U_{test} [66]:

- **Phân chia (Splitting)**: Chọn ngẫu nhiên một số hàng trong ma trận R (một tập con người dùng) vào tập huấn luyện U_{train}, phần còn lại đưa vào tập thực nghiệm U_{test}.
- **Lấy mẫu Bootstrap (Bootstrap sampling)**: Chỉ với một tập dữ liệu ban đầu, thông qua phương pháp lấy mẫu có hoàn lại, ta có thể sinh ra mẫu mới cho vào tập huấn luyện U_{train} và phần còn lại cho vào tập thực nghiệm U_{test}.
- **Kiểm thử chéo (*k*-fold cross validation)**: Chia tập người dùng U thành k tập con có cùng kích cỡ. Tiến hành đánh giá thuật toán tư vấn k lần, tại mỗi lần kiểm nghiệm sử dụng một trong k tập con làm tập thực nghiệm (U_{test}) và các tập con còn lại là tập huấn luyện (U_{train}). Kết quả kiểm nghiệm tổng thể được lấy trung bình từ k lần kiểm nghiệm này.



Hình 1.8. Phương pháp phân chia tập dữ liệu phục vụ cho đánh giá hệ thống tư vấn

Sau khi chia tập người dùng U thành 2 phần U_{train} *và* U_{test}, để đánh giá hệ thống tư vấn hiện thời ta tiến hành như sau: Với mỗi người dùng $u \in U_{test}$, các đánh giá $r_{u,p} \neq 0$ được chia thành hai phần O_u và P_u. O_u được coi là đã biết, trong khi đó P_u là đánh giá cần dự đoán từ dữ liệu huấn luyện U_{train} và O_u (Hình 1.8).

Giả sử phương pháp lọc đưa ra dự đoán cho người dùng trong tập P_u là P'_u . Khi đó kết quả đánh giá cho hệ tư vấn hiện thời sẽ dựa trên sự đối chiếu các đánh giá từ hai tập P_u và P'_u.

Hai nhiệm vụ chính của hệ tư vấn là dự đoán đánh giá và tư vấn danh sách ngắn các sản phẩm cho người dùng hiện thời. Căn cứ theo hai nhiệm vụ đó thì có hai nhóm độ đo đánh giá hệ thống tư vấn tương ứng là: 1) Nhóm độ đo đánh giá độ chính xác của đánh giá dự đoán; 2) Nhóm độ đo đánh giá độ chính xác của danh sách sản phẩm tư vấn. Các độ đo đánh giá hệ tư vấn này đều khai thác thông tin từ P_u và P'_u. Phần trình bày dưới đây sẽ cụ thể hóa các nhóm độ đo này.

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. Có nhiều phương pháp đánh giá sai số phân loại khác nhau đã được đề xuất. Dưới đây là một số độ đo phổ biến.

Độ đo trung bình giá trị tuyệt đối lỗi MAE

Một độ đo phổ biến nhất được sử dụng đánh giá sai số giữa giá trị đánh giá dự đoán và giá trị đánh giá thực tế trong hệ tư vấn là độ đo trung bình giá trị tuyệt đối lỗi MAE (*Mean Absolute Error).*

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

$$MAE_u = \frac{1}{|P_u|} \sum_{x \in P_u} |\hat{r_x}^u - {r_x}^u| \tag{1.17}$$

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

$$MAE = \frac{\sum_{u \in U_{test}} MAE_u}{|U_{test}|} \tag{1.18}$$

Độ đo trung bình lỗi lấy căn RMSE

Một độ đo thông dụng khác cũng được sử dụng để đánh giá giá trị dự đoán là bình phương trung bình lỗi lấy căn *RMSE (Root Mean Square Error)*. $RMSE$ được tính bằng căn bậc hai của trung bình bình phương giữa giá trị thực và giá trị dự đoán.

$$RMSE_u = \sqrt{\frac{1}{|P_u|} \sum_{x \in P_u} (\hat{r_x}^u - {r_x}^u)^2} \tag{1.19}$$

$$RMSE = \frac{\sum_{u \in U_{test}} RMSE_u}{|U_{test}|} \tag{1.20}$$

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

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

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

Giả sử từ ma trận đánh giá R, giá trị đánh giá của người dùng với sản phẩm được chia thành 2 loại: đánh giá thích và đánh giá không thích. Những đánh giá thích là đánh giá có giá trị lớn hơn hoặc bằng một ngưỡng θ cho trước, những giá trị đánh giá không thích là đánh giá có giá trị nhỏ hơn ngưỡng θ.

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn với những sản phẩm thực tế thích bởi người dùng thuộc tập P (Hình 1.8), ta tiến hành xây dựng ma trận nhầm lẫn (Confusion matrix) sau:

Bảng 1.1. Ma trận nhầm lẫn (Confusion matrix)

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

Trong đó :

- Tổng số sản phẩm thuộc P tư vấn cho các người dùng trong U_{test} là $(d + b)$:
  - d sản phẩm tư vấn cũng là những sản phẩm thực tế thích bởi người dùng.
  - b sản phẩm tư vấn nhưng lại là những sản phẩm thực tế không được thích bởi người dùng.
- Tổng số sản phẩm thuộc P không được tư vấn cho các người dùng trong U_{test} là $(a + c)$ gồm :
  - a sản phẩm không được tư vấn và thực tế các sản phẩm này cũng không được thích bởi người dùng.
  - c sản phẩm không được tư vấn nhưng thực tế các sản phẩm này được thích bởi người dùng.

Từ ma trận nhầm lẫn, một số độ đo tính chính xác của danh sách sản phẩm tư vấn được đưa ra, các độ đo này có nguồn gốc từ lĩnh vực học máy. Dưới đây là một số độ đo phổ biến.

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

Hệ tư vấn đưa ra một danh sách ngắn các sản phẩm người dùng có thể thích từ một tập các sản phẩm có sẵn của hệ thống. Điều này tương đương với quá trình lọc thông tin trong lĩnh vực truy vấn thông tin (Information Retrieval – IR). Do vậy, các độ đo hiệu năng tiêu chuẩn trong lọc thông tin thường được sử dụng để đánh giá hiệu năng hệ tư vấn. Hai trong số các độ đo đó là độ chính xác ($Precision$), độ nhạy ($Recall$) [66]. Độ chính xác và độ nhạy được xác định theo công thức sau :

- Độ chính xác ($Precision$)

$$Precision = \frac{d}{b+d} \tag{1.21}$$

Trong đó :

- d sản phẩm tư vấn trong tập P cũng là những sản phẩm thực tế được thích bởi người dùng thuộc U_{test}.
- $b + d$: tổng số sản phẩm trong tập P tư vấn cho người dùng trong U_{test}.

- Độ nhạy ($Recall$)

$$Recall = \frac{d}{c+d} \tag{1.22}$$

Trong đó :

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

Độ chính xác và độ nhạy có giá trị ngược nhau: độ chính xác cao thì độ nhạy thấp và ngược lại. Để cân bằng giữa 2 độ đo này, một độ đo mới được đưa ra đó là *E-measure* [66] theo công thức sau :

$$E - measure = \frac{1}{\alpha\left(\frac{1}{Precision}\right) + (1-\alpha)(\frac{1}{Recall})} \tag{1.23}$$

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

Trong trường hợp $\alpha = 0.5$ khi đó $Precision$ và $Recall$ có vai trò như nhau trong việc đánh giá độ chính xác của hệ thống. Với trường hợp này, độ đo $E - measure$ được định nghĩa với tên mới *F-measure* (hay F1)[66] theo công thức sau:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{1.24}$$

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

Tuy nhiên trên thực tế thì số lượng kết quả trả về sẽ rất lớn mà người dùng không cần thiết lựa chọn hết nên chỉ số $Precision$ không còn mấy ý nghĩa. Vì vậy người ta thường sử dụng chỉ số $Precision@k$ để có thể đánh giá kết quả trả về

chuẩn xác hơn. $Precision@k$ khác với $Precision$ ở chỗ chỉ lựa chọn $top\ k$ sản phẩm có giá trị dự đoán cao nhất để đánh giá.

$$Precision@k = \frac{|P_{true}|}{\min(\mathrm{k}, |\mathrm{P}|)} \tag{1.25}$$

Trong đó:

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

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

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

$$Recall@k = \frac{|P_{true}|}{\min(\mathrm{k}, \mathrm{M})} \tag{1.26}$$

Trong đó:

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

Độ chính xác trung bình tuyệt đối MAP (Mean Average Precision)

MAP là một chỉ số đánh giá phổ biến khác của hệ tư vấn. Ý tưởng của MAP là ngoài vấn đề thể hiện sự hiệu quả của những gợi ý như $Precision$, MAP còn thể hiện tính đúng đắn về thứ hạng của những gợi ý. Chẳng hạn như đối với $Precision$, người dùng chỉ lựa chọn 1 trong 10 gợi ý thì dù đó là gợi thứ 1 hay thứ 10 thì giá trị $Precision$ vẫn là 0,1. Nhưng đối với MAP thì khác, nếu người dùng lựa chọn gợi ý thứ 1 thì giá trị MAP sẽ lớn hơn nếu người dùng chọn gợi ý thứ 10.

Cũng như $Precision$, người ta thường sử dụng $MAP@k$ với k là số lượng sản phẩm hệ thống chọn để tư vấn cho người dùng. $MAP@k$ được định nghĩa thông qua $AP@k$ $(Average\ Precision)$ dưới đây.

$$AP@k = \frac{1}{m}\sum_{i=1}^{k}(Precision@i \text{ nếu sản phẩm thứ } i \text{ phù hợp}) \\ = \frac{1}{m}\sum_{i=1}^{k} Precision@i.rel(i) \tag{1.27}$$

Trong đó:

- $rel(i) = 1$ nếu sản phẩm thứ i phù hợp với người dùng, $rel(i) = 0$ trong trường hợp còn lại.
- m: tổng số lượng sản phẩm liên quan.

Độ đo $AP@k$ được áp dụng để tính độ chính xác trung bình cho mỗi người dùng thuộc tập U_{test}. Trên cơ sở đó, độ chính xác trung bình tuyệt đối $MAP@k$ cho tất cả người dùng trong tập U_{test} được tính bằng trung bình cộng $AP@k$ của các người dùng trong U_{test}.

$$MAP@k = \frac{1}{|U_{test}|}\sum_{i=1}^{|U_{test}|}(AP@k)_{u_i} \tag{1.28}$$

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

Hệ tư vấn không chỉ đơn thuần là một dạng hệ thống thông tin mà nó còn là cả một lĩnh vực nghiên cứu hiện đang rất được các nhà khoa học quan tâm. Kể từ năm 2007 đến nay, hàng năm đều có hội thảo chuyên về hệ tư vấn của ACM (ACM RecSys) cũng như các tiểu ban dành riêng cho RS trong các hội nghị lớn khác như ACM KDD, ACM CIKM,…Ngoài ra, với sự phát triển ngày càng lớn mạnh của hệ tư vấn, một số tập dữ liệu dùng chung (LastFM, MovieLens, Bibsonomy,...) và các phần mềm hỗ trợ/ nguồn mở đã được cung cấp nhằm hỗ trợ các nhà phát triển trong xây dựng sản phẩm cũng như nghiên cứu về hệ tư vấn được nhanh hơn. Bảng dưới đây liệt kê một số phần mềm / nguồn mở hỗ trợ xây dựng và kiểm nghiệm hệ tư vấn được cộng đồng các nhà phát triển cung cấp [66].

Bảng 1.2. Một số phần mềm hỗ trợ nghiên cứu, phát triển hệ tư vấn

Phần mềm	Miêu tả	Ngôn ngữ	URL
Apache Mahout	Thư viện học máy xử lý với các bộ dữ liệu lớn, trong đó bao gồm cài đặt các thuật toán tiên tiến về lọc cộng tác. Việc tích hợp thư viện này vào hệ tư vấn khá ngắn gọn, đơn giản.	Java	http://mahout.apache.org/
Cofi	Thư viện lọc cộng tác cơ sở.	Java	http://www.nongnu.org/cofi/
Crab	Thư viện cung cấp một tập đa dạng các thành phần giúp xây dựng và tinh chỉnh hệ tư vấn từ các thuật toán lọc thông tin truyền thống.	Python	https://github.com/muricoca/crab
Easyrec	Framework phát triển hệ tư vấn trên nền Web.	Java	http://easyrec.org/
LensKit	Thư viện lọc cộng tác cơ sở đưa ra bởi nhóm nghiên cứu GroupLens.	Java	http://lenskit.grouplens.org/
MyMediaLite	Thư viện cài đặt các giải thuật xây dựng và kiểm nghiệm hệ tư vấn cơ sở.	C#/Mono	http://www.mymedialite.net/
PREA	Bộ công cụ cài đặt các thuật toán tư vấn.	Java	http://prea.gatech.edu/
SVDFeature	Thư viện cài đặt giải thuật Matrix Factorization.	C++	http://mloss.org/software/view/333/
Vogoo PHP LIB	Thư viện lọc cộng tác cho	PHP	http://sourceforge.net/

	xây dựng hệ tư vấn trên nền Website.		projects/vogoo/
MLib	Thư viện lọc cộng tác dựa trên mô hình. Trong đó, các người dùng và sản phẩm được mô tả bằng tập các nhân tố tiềm ẩn (Latent factors). Mlib sử dụng thuật toán ALS (Alternating Least Squares) để học các dữ liệu tiềm ẩn này.	Java	http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html
LibRec	Thư viện của Java GPL-licensed dùng để cài đặt hệ tư vấn. Thư viện này cài đặt sẵn các thuật toán tư vấn tiên tiến hiện nay (khoảng hơn 70 thuật toán).	Java (phiên bản 1.7 trở lên)	http://www.librec.net/
RecommenderLab	Package cung cấp cơ sở hạ tầng để phát triển và kiểm thử các thuật toán tư vấn trong một framework thống nhất, xử lý và phân tích dữ liệu nhanh chóng.	R	https://github.com/mhahsler/recommenderlab

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. 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 sản phẩm phù hợp với người dùng hiện

thời. Ba giai đoạn này kết hợp với nhau tạo thành một qui trình xây dựng hệ tư vấn nói chung.

Có nhiều đề xuất khác nhau để giải quyết bài toán tư vấn theo "Qui trình xây dựng hệ tư vấn". Tuy nhiên về cơ bản thì hệ tư vấn được chia thành hai hướng tiếp cận tùy vào việc lựa chọn loại thông tin, mô hình học và dự đoán sản phẩm mới cho người dùng, đó là: 1) Hệ tư vấn với cách tiếp cận truyền thống; 2) Hệ tư vấn mở rộng cách tiếp cận truyền thống. Trong đó hướng tiếp cận thứ 2 được mở rộng ra từ hướng tiếp cận 1 trong xu thế gia tăng đa dạng các nguồn thông tin thu thập đa chiều và nghiên cứu cải tiến các phương pháp lọc tin truyền thống trong hệ tư vấn. Theo cả hai hướng tiếp cận này thì chất lượng của hệ tư vấn sẽ phụ thuộc vào hiệu quả của phương pháp lọc thông tin trên các thông tin đầu vào của hệ thống. Bên cạnh những nghiên cứu cơ bản và mở rộng về hệ tư vấn về lý thuyết, luận án cũng tiến hành khảo sát các phương pháp và độ đo đánh giá hệ tư vấn, đây là cơ sở để các nhà phát triển lựa chọn một mô hình tư vấn phù hợp cho hệ thống của mình. Ngoài ra luận án cũng chỉ ra các nguồn tài nguyên hỗ trợ học tập, nghiên cứu hệ tư vấn phổ biến hiện nay.

Các nội dung trên lần lượt được trình bày và xâu chuỗi lại trong các phân mục của chương 1- Tổng quan về hệ tư vấn. Trên cơ sở đó, luận án tập trung vào nghiên cứu phát triển một số phương pháp tư vấn với mục tiêu cụ thể sau:

- Nghiên cứu và đề xuất phương pháp hạn chế ảnh hưởng vấn đề dữ liệu thưa của hệ tư vấn cộng tác truyền thống dựa trên mô hình đồ thị và mở rộng cho phát triển hệ tư vấn cộng tác theo ngữ cảnh. Phương pháp đề xuất được trình bày trong Chương 2.
- Nghiên cứu và đề xuất phương pháp lọc kết hợp bằng đồng huấn luyện để nâng cao chất lượng tư vấn. Phương pháp đề xuất được trình bày trong Chương 3.

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

Nội dung chương 2 trình bày kết quả nghiên cứu của luận án về đề xuất phương pháp lọc cộng tác theo ngữ cảnh dựa trên mô hình đồ thị, nhằm tích hợp hiệu quả thông tin ngữ cảnh và giải quyết vấn đề dữ liệu thưa cho hệ tư vấn theo ngữ cảnh. Trước hết, vấn đề dữ liệu thưa của lọc cộng tác và các hướng giải quyết được trình bày trong Mục 2.1, trong đó tập trung vào hướng tiếp cận liên quan trực tiếp tới đề xuất của luận án. Mục 2.2 trình bày đề xuất về độ đo tương tự mới cho lọc cộng tác dựa trên mô hình đồ thị (Khi không có ngữ cảnh). Mục 2.3 trình bày bài toán tư vấn theo ngữ cảnh kèm các hướng tiếp cận giải quyết, trên cơ sở đó tác giả đưa ra đề xuất phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh. Mục 2.4 trình bày về kết quả thực nghiệm, so sánh và đánh giá phương pháp đề xuất trong sự so sánh với các phương pháp xây dựng hệ tư vấn theo ngữ cảnh cơ sở. Mục cuối cùng 2.5 là kết luận và hướng nghiên cứu tiếp theo. Đây cũng là những kết quả nghiên cứu đã được công bố trong [C1][C3][C7][C4][J2].

2.1. Đặt vấn đề

Như đã trình bày trong mục 1.5.1.3 của Chương 1, một trong số khó khăn chính mà các phương pháp lọc cộng tác gặp phải là vấn đề dữ liệu thưa [1][10]. Vấn đề dữ liệu thưa ảnh hưởng trực tiếp đến kết quả tính toán mức độ tương tự, xác định tập láng giềng và nhiều vấn đề liên quan khác trong lọc cộng tác. Chính vì vậy, hạn chế ảnh hưởng vấn đề dữ liệu thưa là một trong những trọng tâm nghiên cứu của lọc cộng tác. Để giải quyết vấn đề dữ liệu thưa cho lọc cộng tác, một số hướng tiếp cận được đưa ra, điển hình như 2 hướng: 1) Giảm số chiều của ma trận đánh giá; 2) Khai thác các mối liên hệ gián tiếp trên ma trận đánh giá.

Hướng tiếp cận giảm số chiều của ma trận đánh giá được thực hiện bằng cách tạo nên ma trận tương tác đặc hơn so với ma trận thưa thớt đánh giá ban đầu, sau đó

sử dụng ma trận này để tính toán mức độ tương quan giữa người dùng hoặc sản phẩm. Chiến lược đơn giản nhất để giảm số chiều của ma trận đánh giá là tạo lập nên các cụm sản phẩm, cụm người dùng hoặc cụm dữ liệu chứa cả người dùng và sản phẩm, sau đó sử dụng những cụm này như những đơn vị cơ bản để sinh ra dự đoán [28][67]. Ngoài ra, phương pháp giảm số chiều của ma trận đánh giá bằng các kỹ thuật thống kê là một chiến lược khác cũng rất được quan tâm nghiên cứu, như việc sử dụng phương pháp phát hiện ngữ nghĩa ẩn (LSM) dựa trên kỹ thuật phân rã giá trị riêng (SVD), cải tiến phương pháp phân cụm sử dụng kỹ thuật phân tích thành phần chính (PCA)… [68][69][70][71]. Tuy nhiên, trong nhiều trường hợp thông tin hữu ích có thể bị mất trong quá trình giảm chiều ma trận làm giảm hiệu quả tư vấn.

Hướng tiếp cận chính thứ 2 giải quyết vấn đề dữ liệu thưa bằng việc khai thác các mối liên hệ gián tiếp trên ma trận đánh giá, trong đó mô hình đồ thị là dạng biểu diễn tất cả các mối liên kết tiêu biểu. Phương pháp lọc cộng tác dựa trên mô hình đồ thị đề xuất đầu tiên bởi Huang và các cộng sự [35] được thực hiện căn cứ vào việc tính mức độ phù hợp của sản phẩm với người dùng thông qua các mối liên kết từ đỉnh người dùng tới đỉnh sản phẩm trên đồ thị. Phương pháp Random Walk được đề xuất bởi Fouss [72] căn cứ trên mức độ tương tự giữa các đỉnh trên đồ thị. Một số hướng tiếp cận khác coi việc dự đoán mức độ phù hợp của người dùng với sản phẩm là việc dự đoán liên kết sử dụng các phương pháp học máy có giám sát [73]. Yang và Toni [74] đề xuất phương pháp giảm số chiều của dữ liệu dùng cho hệ tư vấn bằng việc phân cụm các đồ thị người dùng. Deladiennee và Naudet [75] đưa ra giải pháp biểu diễn tri thức lĩnh vực để nâng cao chất lượng tư vấn trong các lĩnh vực phức tạp. Ngoài ra một số nghiên cứu cũng mở rộng tích hợp thêm các thông tin về đặc trưng sản phẩm cùng với thông tin người dùng, sản phẩm trên một mô hình đồ thị để phục vụ cho hệ tư vấn dựa vào phương pháp lọc kết hợp [54]. Trong đó có thể nói công trình nghiên cứu của Huang và các cộng sự [35] được coi là đặt nền móng cho ý tưởng xây dựng hệ tư vấn dựa trên mô hình đồ thị ở qui mô tổng quát.

Mô hình đề xuất bởi Huang và các cộng sự [35] cho phép biểu diễn tất cả quan điểm của người dùng đối với các sản phẩm trong lọc cộng tác bằng một đồ thị hai phía (Bipart Graph Model), trong đó một phía là tập đỉnh người dùng, phía còn lại là tập đỉnh sản phẩm, mỗi cạnh nối từ đỉnh người dùng tới đỉnh sản phẩm được thiết lập nếu người dùng đã đưa ra phản hồi tới sản phẩm (đánh giá / mua…). Trên cơ sở biểu diễn đồ thị đó, Huang và các cộng sự khai thác tất cả các mối quan hệ bắc cầu từ đỉnh người dùng tới đỉnh sản phẩm trên đồ thị để thực hiện tính toán trực tiếp mức độ phù hợp của các sản phẩm mới với người dùng hiện thời. Việc tính toán mức độ phù hợp của các sản phẩm mới với người dùng hiện thời sẽ được tính bằng tổng trọng số các đường đi từ đỉnh người dùng tới đỉnh sản phẩm trên đồ thị, trong đó trọng số mỗi đường đi có giá trị bằng tích trọng số các cạnh trên đường đi đó. Từ đó hệ thống sẽ chọn ra K sản phẩm mới có mức độ phù hợp cao nhất để tư vấn cho người dùng hiện thời.

Phương pháp của Huang và các cộng sự được đánh giá là cải thiện chất lượng tư vấn hơn các phương pháp lọc cộng tác dựa trên độ tương quan trước đó (Đề cập trong 1.5.1). Tuy nhiên theo cách tiếp cận này, một số mối quan hệ không cần thiết hoặc gây nhiễu có thể xuất hiện trong khi mở rộng độ dài đường đi từ đỉnh người dùng tới đỉnh sản phẩm trên đồ thị, điều này hiện đang ảnh hưởng trực tiếp tới việc tính mức độ phù hợp của người dùng với sản phẩm, dẫn tới làm giảm độ chính xác tư vấn.

Nhằm phát huy thế mạnh của mô hình đồ thị trong việc 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 trong phương pháp của Huang và các cộng sự nói trên, trong Mục 2.2 tiếp theo 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. 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 mở rộng độ đo tương tự này cho phát triển hệ tư vấn cộng tác theo ngữ cảnh trong mục 2.3.

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

Về bản chất, các phương pháp lọc cộng tác dựa trên mô hình đồ thị đều tiến hành giải quyết ba vấn đề: *1) Phương pháp biểu diễn đồ thị cho lọc cộng tác; 2) Phương pháp tính toán mức độ tương tự giữa các cặp người dùng (hoặc sản phẩm) hoặc tính mức độ phù hợp của sản phẩm với người dùng; 3) Xây dựng thuật toán dự đoán quan điểm của người dùng đối với các sản phẩm dựa vào đồ thị*.

Trong phần này, tác giả mở rộng mô hình biểu diễn trên đồ thị được đề xuất bởi Huang và các cộng sự nhằm đề xuất phương pháp tính mức độ tương tự giữa các cặp người dùng hoặc giữa các cặp sản phẩm cho hệ tư vấn cộng tác dựa trên biểu diễn đồ thị này. Chi tiết của phương pháp tính mức độ tương tự cho lọc cộng tác dựa trên mô hình đồ thị đề xuất được trình bày trong hai nội dung dưới đây: 1) Biểu diễn đồ thị cho lọc cộng tác; 2) Độ đo tương tự cho lọc cộng tác dựa trên biểu diễn đồ thị.

2.2.1. Biểu diễn đồ thị cho lọc cộng tác

Giả sử cho hệ lọc cộng tác gồm N người dùng $U = \{u_1, u_1, \dots, u_N\}$ và M sản phẩm $P = \{p_1, p_2, \dots, p_M\}$. Mối quan hệ giữa tập người dùng U và tập sản phẩm P được biểu diễn thông qua ma trận đánh giá là $R = [r_{ij}]$ với $i = 1, 2, \dots, N$; $j = 1, 2, \dots, M$.

Không hạn chế tính tổng quát của bài toán, ta giả sử $r_{ij} = v$ nếu người dùng $u_i \in U$ đánh giá sản phẩm $p_j \in P$ ở mức độ v, trong đó $v \in (0, 1]$ (Giá trị v càng cao thì mức độ ưa thích càng lớn) và $r_{ij} = 0$ trong trường hợp còn lại.

$$r_{ij} = \begin{cases} v & Nếu\ người\ dùng\ u_i\ thích\ sản\ phẩm\ p_j\ ở\ mức\ độ\ v. \\ 0 & Nếu\ người\ dùng\ u_i\ chưa\ biết\ hoặc\ chưa\ đánh\ giá\ sản\ phẩm\ p_j. \end{cases} \quad (2.1)$$

Biểu diễn ma trận đánh giá theo (2.1) sẽ không ảnh hưởng đến các hệ thống lọc cộng tác sử dụng đánh giá nhị phân {0,1} hoặc có nhiều mức đánh giá trong khoảng [0,1]. Đối với các bộ dữ liệu có giá trị đánh giá $r_{ij} \in \{1, 2, .., g\}$, ta chỉ cần thực hiện phép biến đổi đơn giản chuyển $r_{ij} = \frac{r_{ij}}{g}$. Mục đích của việc chuyển đổi

$r_{ij} \in [0,1]$ để sử dụng trong phương pháp tính toán mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm. Phép biến đổi này vẫn bảo toàn được mức độ đánh giá theo thứ tự khác nhau của các hệ lọc cộng tác. Đây là một biểu diễn mở rộng của Huang và các cộng sự đã thực hiện trong [35].

Ví dụ với hệ lọc cộng tác được cho trong Bảng 2.1 gồm 3 người dùng $U = \{u_1, u_2, u_3\}$ và 4 sản phẩm $P = \{p_1, p_2, p_3, p_4\}$. Mỗi người dùng đưa ra các đánh giá của mình về các sản phẩm theo thang bậc {1, 2, 3, 4, 5}. Giá trị $r_{ij} = 0$ được hiểu là người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j. Giá trị r_{31} =? là giá trị đánh giá hệ thống cần dự đoán cho người dùng u_3 với sản phẩm p_1. Như vậy hệ lọc cộng tác được cho trong Bảng 2.1 sẽ được chuyển đổi biểu diễn theo (2.1) thành Bảng 2.2.

Bảng 2.1. Ví dụ ma trận đánh giá của lọc cộng tác

	p_1	p_2	p_3	p_4
u_1	5	0	4	0
u_2	0	3	4	0
u_3	?	3	?	2

Bảng 2.2. Ma trận đánh giá chuyển đổi

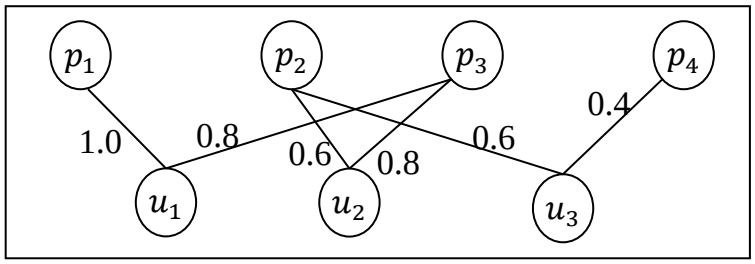
	p_1	p_2	p_3	p_4
u_1	1	0	0.8	0
u_2	0	0.6	0.8	0
u_3	?	0.6	?	0.4

Hệ lọc cộng tác với ma trận đánh giá xác định theo (2.1) 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).

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

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

Ví dụ, với hệ lọc cộng tác được cho trong Bảng 2.2, khi đó đồ thị hai phía biểu diễn cho lọc cộng tác được thể hiện trong Hình 2.1.



Hình 2.1. Đồ thị biểu diễn cho lọc cộng tác

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

Như đã đề cập trong Mục 2.1, các độ đo tương quan tính toán mức độ tương tự giữa các cặp người dùng $\{u_i, u_j\}$ dựa trên tập các sản phẩm cả hai người dùng cùng đánh giá $P_{ij} \neq \emptyset$. Việc làm này được thực hiện trên đồ thị biểu diễn cho lọc cộng tác G bằng cách tính toán tổng trọng số của tất cả các đường đi có độ dài 2 từ đỉnh $u_i \in U$ đến đỉnh $u_j \in U$. Trọng số của mỗi đường đi độ dài 2 từ đỉnh $u_i \in U$ đến đỉnh $u_j \in U$ được tính bằng tích trọng số của các cạnh tương ứng. Ví dụ để tính toán mức độ tương tự giữa người dùng u_1 và u_2 Hình 2.1, ta tính tổng trọng số các đường đi độ dài 2 từ đỉnh u_1 đến đỉnh u_2. Giữa u_1 và u_2 chỉ có duy nhất một đường đi độ dài 2: $u_1 - p_3 - u_2$, trọng số của đường đi này được tính là 0.8 * 0.8 = 0.64. Tương tự, mức độ tương tự giữa người dùng u_2 và u_3 là 0.6*0.6 = 0.36. Như vậy, phương pháp tính toán mức độ tương tự giữa 2 người dùng dựa trên các độ đo tương quan được xem như việc tính toán tổng trọng số các đường đi độ dài 2 giữa 2 đỉnh người dùng đó trên đồ thị $G =< V, E >$.

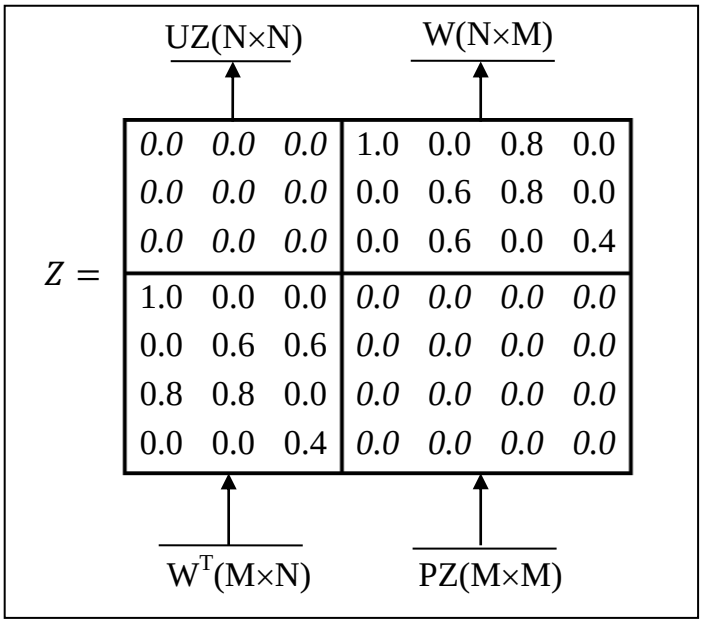
Khi tập các sản phẩm cả hai người dùng cùng đánh giá $P_{ij} = \emptyset$, các độ đo tương quan sẽ không xác định được mức độ tương tự giữa người dùng u_i và người dùng u_j [10][14]. Ví dụ với người dùng u_1 và u_3 trong Bảng 2.2 sẽ không xác định được giá trị tương tự bằng các độ đo tương quan. Nguyên nhân chính để người dùng u_i và u_j không xác định được mức độ tương tự bằng các độ tương quan vì ma trận đánh giá quá thưa. Các giá trị đánh giá $r_{ij} = 0$ trên bộ dữ liệu MovieLens là 98.7%, trên bộ dữ liệu BookCrossing là 99.1%. Tuy nhiên, quan sát những mối quan hệ khác giữa các cặp người dùng trên đồ thị ta có thể thấy giữa họ vẫn tồn tại một mức độ tương tự tiềm ẩn nào đó. Ví dụ, dựa vào biểu diễn đồ thị Hình 2.1, u_1 và u_3 đều tương tự với u_2 và do vậy có thể tương tự nhau. Khai thác được những mối quan hệ này sẽ cải thiện đáng kể chất lượng dự đoán của các phương pháp User-Based k-NN và hạn chế được vấn đề dữ liệu thưa của lọc cộng tác.

Để khai thác được những mối quan hệ gián tiếp trên đồ thị, tác giả thực hiện mở rộng độ dài đường đi giữa 2 đỉnh người dùng trên đồ thị. Ví dụ giữa u_1 và u_3 tồn tại đường đi độ dài *4*: $u_1 - p_3 - u_2 - p_2 - u_3$, trọng số của đường đi này là 0.8*0.8*0.6*0.6=0.2304. Vì G là đồ thị hai phía nên độ dài đường đi giữa 2 đỉnh người dùng luôn là một số chẵn (2, 4, 6, 8,...). Mặt khác, trọng số mỗi cạnh của đồ thị là một số dương nhỏ hơn 1 nên các đường đi có độ dài lớn sẽ được đánh trọng số thấp, đường đi có độ dài nhỏ sẽ được đánh trọng số cao. 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. Khi đó, mức độ tương tự giữa các cặp người dùng được xác định căn cứ vào quá trình tìm kiếm trên đồ thị tất cả các đường đi có độ dài L từ đỉnh người dùng đến đỉnh người dùng. Với cách tiếp cận này, mức độ tương tự giữa các cặp người dùng được xác định dựa trên tất cả các mối quan hệ trực tiếp hoặc gián tiếp trên đồ thị G. Việc làm này được xác định thông qua ma trận trọng số tổng quát biểu diễn đồ thị G dưới đây.

Gọi $Z = \{z_{ij}\}$ là ma trận trọng số biểu diễn đồ thị $G(i = 1,2, \dots (N + M);\ j = 1,2, \dots (N + M))$. Khi đó ma trận vuông Z sẽ được chia thành bốn phần theo công thức (2.4). Trong đó, ma trận vuông $UZ(N \times N)$ biểu diễn mối quan hệ giữa người dùng và người dùng, $PZ(H \times H)$ biểu diễn mối quan hệ giữa sản phẩm với sản phẩm, $W(N \times M)$ được xác định theo (2.3) biểu diễn mối quan hệ giữa người dùng và sản phẩm, $W^T(M \times N)$ là chuyển vị của $W(N \times M)$ biểu diễn mối quan hệ giữa sản phẩm và người dùng. Các phần tử của ma trận $UZ(N \times N)$, $PZ(M \times M)$ ban đầu đều có giá trị 0, tương ứng với mức độ tương tự giữa các cặp người dùng hoặc giữa các cặp sản phẩm không xác định tại thời điểm ban đầu.

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

Ví dụ, với hệ lọc cộng tác được cho trong Bảng 2.2, đồ thị hai phía biểu diễn cho lọc cộng tác được thể hiện trong Hình 2.1, khi đó các thành phần của ma trận trọng số được thể hiện trong Hình 2.2 sau.



Hình 2.2. Ma trận trọng số biểu diễn đồ thị hai phía G

Như vậy, với việc biểu diễn đồ thị G dưới dạng ma trận trọng số Z cho phép thể hiện đầy đủ mối quan hệ giữa các người dùng, các sản phẩm, người dùng – sản

phẩm. Ma trận Z cũng thể hiện rõ ràng sự liên quan theo các chiều của các ma trận con.

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:

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

Trong đó, uz_{ij}^L là mức độ tương tự giữa đỉnh người dùng u_i và đỉnh người dùng u_j căn cứ trên mô hình đồ thị G, giá trị này phụ thuộc vào độ dài đường đi L từ đỉnh u_i tới đỉnh u_j trên đồ thị. Do vậy, một vấn đề đặt ra là với mỗi người dùng $u_i \in U$ giá trị của L được lấy bằng bao nhiêu cho tốt. 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 được xác định theo (2.5).*

Chứng minh. Giả sử đồ thị biểu diễn cho các hệ lọc cộng tác $G =< V, E >$ liên thông. Khi đó luôn tồn tại đường đi từ đỉnh $u_i \in U$ đến mọi $u_j \in U$ trên đồ thị. Vì $G =< V, E >$ là đồ thị hai phía được biểu diễn theo (2.4), nên luôn tồn tại số chẵn L sao cho từ $u_i \in U$ đến $u_j \in U$ được nối bằng đúng L cạnh. Do uz_{ij}^L được xác định theo (2.5) là tổng trọng số các đường đi có độ dài L; Trọng số mỗi đường đi có độ dài L là tích của trọng số các cạnh có $w_{ij}^L \neq 0$, vì vậy nên $uz_{ij}^L \neq 0$ là điều cần chứng minh.

Như vậy, với mỗi người dùng $u_i \in U$ cần được xác định mức độ tương tự với các người dùng $u_j \in \{U \backslash u_i\}$, ta chỉ cần chọn giá trị L *nhỏ nhất* để $uz_{ij}^L \neq 0$. Ví dụ với hệ lọc cộng tác được biểu diễn bằng ma trận trọng số Z trên Hình 2.2, ta tính toán được $UZ^2(3\times3), UZ^4(3\times3), UZ^6(3\times3)$ theo (2.5). Dựa vào đó ta xác định được

$L = 2$ cho người dùng u_2, $L = 4$ cho người dùng u_1 và u_3 và không cần thực hiện với giá trị $L = 6$.

$$UZ^2 = \begin{bmatrix} 1.64 & 0.64 & 0.00 \\ 0.64 & 1.00 & 0.36 \\ 0.00 & 0.36 & 0.52 \end{bmatrix} \quad UZ^4 = \begin{bmatrix} 3.0992 & 1.6896 & 0.2304 \\ 1.6896 & 1.5392 & 0.5472 \\ 0.2304 & 0.5472 & 0.4000 \end{bmatrix}$$

$$UZ^6 = \begin{bmatrix} 6.164032 & 3.756032 & 0.728064 \\ 3.756032 & 2.817536 & 0.838656 \\ 0.728064 & 0.838656 & 0.404992 \end{bmatrix}$$

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

Lập luận tương tự mục 2.2.2.1 ở trên cho việc tính toán mức độ tương tự giữa các cặp sản phẩm dựa trên biểu diễn đồ thị cho lọc cộng tác $G =< V, E >$. Khi đó, mức độ tương tự giữa sản phẩm $p_x \in P$ và sản phẩm $p_y \in P$ được ước lượng bằng tổng trọng số của tất cả các đường đi độ dài L đi từ đỉnh p_x đến đỉnh p_y trên đồ thị, với trọng số của mỗi đường đi được tính bằng tích trọng số các cạnh tương ứng. Vì G là đồ thị hai phía nên độ dài đường đi từ giữa 2 đỉnh sản phẩm luôn là một số chẵn (2, 4, 6, 8,...). Bằng cách tiếp cận này, mức độ tương tự giữa các cặp sản phẩm được xác định dựa trên tất cả các mối quan hệ trực tiếp hoặc gián tiếp trên đồ thị G.

Việc làm này được xác định dựa vào ma trận trọng số tổng quát Z biểu diễn đồ thị G (Hình 2.2). Khi đó, mức độ tương tự giữa các cặp sản phẩm được tính toán theo công thức (2.6) sau:

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

Mức độ tương tự giữa các cặp sản phẩm xác định theo (2.6) cũng phụ thuộc vào độ dài đường đi L từ đỉnh sản phẩm đến đỉnh sản phẩm trên đồ thị. Do vậy, với mỗi sản phẩm $p_x \in P$ ta cũng cần xác định giá trị của L để thực hiện tính toán. Định lý 2.2 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 2.2. *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} *được xác định theo (2.6).*

Việc chứng minh Định lý 2.2 cũng được lập luận tương tự như Định lý 2.1. Kết quả này cho phép ta xác định giá trị L *nhỏ nhất* để $pz_{xy}^{L} \neq 0$ với mọi $p_x, p_y \in P$. Ví dụ với hệ lọc cộng tác được biểu diễn bằng ma trận trọng số Z trên Hình 2.2, ta tính toán được $PZ^2(4\times4), PZ^4(4\times4), PZ^6(4\times4)$ theo (2.6). Dựa vào đó ta xác định được $L = 4$ đối với sản phẩm p_2 và p_3 , $L = 6$ đối với sản phẩm p_1 và p_4.

$$PZ^2 = \begin{bmatrix} 1.00 & 0.00 & 0.80 & 0.00 \\ 0.00 & 0.72 & 0.48 & 0.24 \\ 0.80 & 0.48 & 1.28 & 0.00 \\ 0.00 & 0.24 & 0.00 & 0.16 \end{bmatrix} \qquad PZ^4 = \begin{bmatrix} 1.6400 & 0.3840 & 1.8240 & 0.0000 \\ 0.3840 & 0.8064 & 0.9600 & 0.2112 \\ 1.8240 & 0.9600 & 2.5088 & 0.1152 \\ 0.0000 & 0.2112 & 0.1152 & 0.0832 \end{bmatrix}$$

$$PZ^6 = \begin{bmatrix} 3.099200 & 1.152000 & 3.831040 & 0.092160 \\ 1.152000 & 1.092096 & 1.923072 & 0.227328 \\ 3.831040 & 1.923072 & 5.131265 & 0.248832 \\ 0.092160 & 0.227328 & 0.248832 & 0.064000 \end{bmatrix}$$

Độ đo tương tự đề xuất trong công thức (2.5) và (2.6) cho lọc cộng tác dựa trên biểu diễn đồ thị cho phép khai thác đầy đủ mối quan hệ trực tiếp và bắc cầu giữa các cặp đỉnh người dùng và các cặp đỉnh sản phẩm trên đồ thị vào quá trình dự đoán và tư vấn, từ đó giúp hạn chế ảnh hưởng của vấn đề dữ liệu thưa của lọc cộng tác. Trên cơ sở đó, luận án tiến hành mở rộng độ đo tương tự nêu trên cho phát triển hệ tư vấn cộng tác theo ngữ cảnh trong nội dung 2.3 dưới đây.

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

Hệ tư vấn với cách tiếp cận truyền thống chỉ quan tâm tới những đối tượng chính là người dùng, sản phẩm và phản hồi của người dùng với sản phẩm, chứ không quan tâm đến các thông tin ngữ cảnh bên ngoài có tác động đến quyết định của người dùng hay không. Tuy nhiên trên thực tế, sở thích của người dùng lại

không cố định. Ví dụ một người trời nóng thì thích ăn kem, uống sinh tố, nhưng khi trời lạnh lại thích ăn phở, uống cà phê nóng. Hoặc cùng một bộ phim nhưng trời mưa thì thích xem còn trời khô ráo thì có khi lại không thích. Có thể nói sở thích của người dùng bị tác động nhiều bởi những yếu tố ngữ cảnh bên ngoài. Do vậy việc xem xét kết hợp ngữ cảnh vào các hệ thống tư vấn là một chủ đề đang rất được quan tâm nghiên cứu trong những năm gần đây. Các hệ tư vấn theo ngữ cảnh được chứng minh là cải thiện đáng kể chất lượng tư vấn trong nhiều ứng dụng thực tế so với hệ tư vấn với cách tiếp cận truyền thống trước đây, qua đó giúp tăng độ tin cậy của người dùng với hệ thống [6][76] . Ví dụ một số hệ tư vấn dựa trên ngữ cảnh như hệ tư vấn phim và âm nhạc [77], truyền thông xã hội [78][79][80] tư vấn địa điểm yêu thích [81], thương mại điện tử [82], tư vấn dịch vụ Web [83], sức khỏe [84], dịch vụ học tập trực tuyến [85], dịch vụ giao thông [86]. Tầm quan trọng của ngữ cảnh cũng được khẳng định khi đưa ra quyết định trong các hệ thống IoT (Internet of Things) [87].

Trong nội dung dưới đây, tác giả sẽ trình bày về yếu tố ngữ cảnh (context) trong hệ tư vấn và giới thiệu khái quát những phương pháp kết hợp yếu tố ngữ cảnh vào một hệ tư vấn, cùng với những hạn chế còn tồn tại của mỗi phương pháp. Từ đó tác giả đề xuất một giải pháp xây dựng hệ tư vấn cộng tác theo ngữ cảnh trên cơ sở độ đo tương tự cho lọc cộng tác dựa trên mô hình đồ thị đã được trình bày trong Mục 2.2. Giải pháp đưa ra được chứng minh là cho phép tích hợp đầy đủ thông tin ngữ cảnh và hạn chế vấn đề dữ liệu thưa của lọc cộng tác trong sự so sánh với các phương pháp tư vấn không có ngữ cảnh và có ngữ cảnh cơ sở.

2.3.1. Ngữ cảnh

Định nghĩa ngữ cảnh

Theo như [6] : “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”.

Phân loại ngữ cảnh

Hai hướng tiếp cận chính trong phân loại ngữ cảnh dùng trong hệ tư vấn: 1) Phân loại ngữ cảnh theo cách thức thu thập và mức độ thay đổi thông tin ngữ cảnh trong hệ tư vấn; 2) Phân loại ngữ cảnh theo các loại thông tin ngữ cảnh phổ biến sử dụng trong các ứng dụng tư vấn.

Với hướng tiếp cận 1, việc phân loại ngữ cảnh sử dụng sản phẩm của người dùng trong hệ tư vấn (Bảng 2.3) được căn cứ trên hai tiêu chí [6] sau:

a) Mức độ thu thập các thông tin ngữ cảnh trong hệ tư vấn (Đầy đủ, một phần hoặc không xác định). Các thông tin ngữ cảnh sử dụng cho hệ tư vấn có thể được thu thập tường minh căn cứ trên ngữ cảnh sử dụng sản phẩm do người dùng nhập vào hoặc ngầm định thông qua tương tác của người dùng với hệ thống và từ các thiết bị, cảm biến liên quan kết nối trong hệ thống (Thời gian, thời tiết, nơi chốn…).
b) Mức độ thay đổi thông tin ngữ cảnh theo thời gian (Không thay đổi hoặc thay đổi).

Bảng 2.3. Phân loại ngữ cảnh thu thập được cho hệ tư vấn

Mức độ thay đổi thông tin ngữ cảnh theo thời gian	**Mức độ thu thập các thông tin ngữ cảnh**		
	Đầy đủ	Một phần	Không xác định
Không thay đổi	Thông tin ngữ cảnh được thu thập đầy đủ và bất biến.	Thông tin ngữ cảnh được thu thập một phần và bất biến.	Thông tin ngữ cảnh có thể được suy luận ngầm định từ dữ liệu của hệ thống.
Thay đổi	Thông tin ngữ cảnh được thu thập đầy đủ và thường xuyên thay đổi.	Thông tin ngữ cảnh được thu thập một phần và thường xuyên thay đổi.	Không thu thập được thông tin ngữ cảnh.

Với hướng tiếp cận 2, có 4 loại thông tin ngữ cảnh phổ biến được người dùng sử dụng khi đưa ra đánh giá cho các sản phẩm trong các hệ tư vấn [6], đó là:

- Ngữ cảnh vật lý (Physical context): Các ngữ cảnh thuộc loại này như thời gian, địa điểm , thời tiết, ánh sáng, nhiệt độ.
- Ngữ cảnh xã hội (Social context), như: Thông tin về mối quan hệ giữa người dùng với các người dùng hoặc các nhóm liên quan.
- Ngữ cảnh phương tiện tương tác (Interaction media context), như: Những thiết bị sử dụng truy cập vào hệ thống (điện thoại di động, cảm biến,…), những loại dữ liệu đa phương tiện được duyệt và cá nhân hóa (văn bản, âm thanh, hình ảnh,..).
- Ngữ cảnh tâm lý (Model context) thể hiện tâm lý của người dùng, như: Mục tiêu người dùng, tâm trạng, trải nghiệm và khả năng nhận thức của người dùng.

2.3.2. Bài toán tư vấn theo ngữ cảnh

Đối với hệ tư vấn truyền thống, ta chỉ quan tâm tới mối quan hệ giữa hai nhóm đối tượng là người dùng và sản phẩm để đưa ra dự đoán. Khi đó bài toán tư vấn truyền thống được biểu diễn dựa trên ma trận đánh giá hai chiều sau:

$$R_o: U \times P \rightarrow R \tag{2.7}$$

Trong khi đó, đối với hệ tư vấn theo ngữ cảnh, ngoài thông tin về hai đối tượng người dùng và sản phẩm, hệ thống còn quan tâm tới những yếu tố ngữ cảnh khi người dùng đánh giá một sản phẩm để đưa ra dự đoán. Khi đó bài toán tư vấn theo ngữ cảnh sẽ dựa trên ma trận đánh giá đa chiều (Multi-dimensional matrix) như sau:

$$R_1: U \times P \times C \rightarrow R \tag{2.8}$$

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.

Ví dụ cho Bảng 2.4 là ma trận đánh giá đa chiều của hệ tư vấn cộng tác theo ngữ cảnh, gồm 3 người dùng $U = \{u_1, u_2, u_3\}$, 2 sản phẩm $P = \{p_1, p_2\}$, kèm thông tin về các chiều ngữ cảnh.

Một số thuật ngữ qui ước được sử dụng trong các hệ tư vấn dựa vào ngữ cảnh được biết đến đó là: *Chiều ngữ cảnh (Context Dimension), điều kiện ngữ cảnh (Context Condition), tình huống ngữ cảnh (Context Situation)* [6][88].

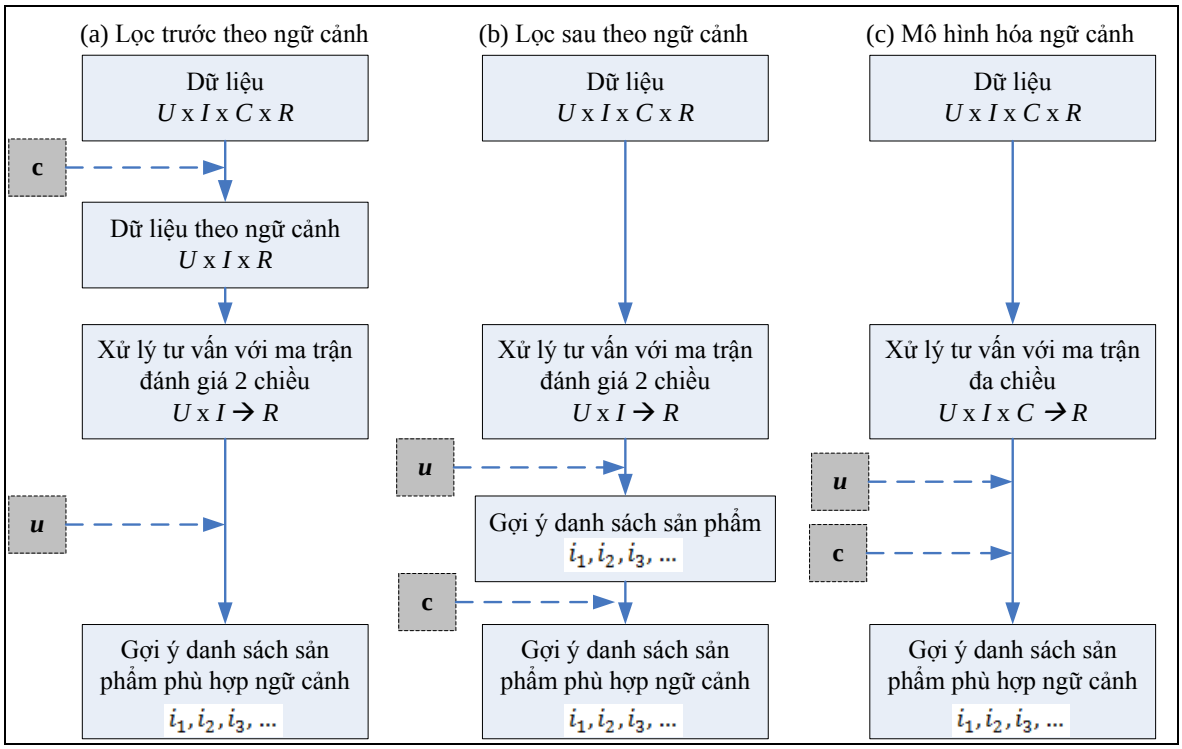
- *Chiều ngữ cảnh* hay còn được biết đến là các biến ngữ cảnh, ví dụ: *"Thời gian", "Địa điểm", "Bạn đồng hành".*
- *Điều kiện ngữ cảnh* là một giá trị của một chiều ngữ cảnh được đánh giá bởi người dùng. Ví dụ: Chiều ngữ cảnh *"Thời gian"* có thể nhận 1 trong 2 điều kiện ngữ cảnh ("Cuối tuần","Trong tuần"), chiều ngữ cảnh *"Địa điểm"* có 2 điều kiện ngữ cảnh ("Tại nhà", "Tại rạp"), chiều ngữ cảnh *"Bạn đồng hành"* có 3 điều kiện ngữ cảnh ("Trẻ em","Gia đình","Đối tác").
- *Tình huống ngữ cảnh* là thuật ngữ dùng để chỉ một bộ giá trị điều kiện ngữ cảnh của các chiều ngữ cảnh tương ứng được đánh giá bởi người dùng, ví dụ người dùng u_1 đánh giá 5 cho sản phẩm p_1 trong tình huống ngữ cảnh ("Cuối tuần","Tại nhà","Trẻ em").

Bảng 2.4. Ma trận đánh giá đa chiều của lọc cộng tác theo ngữ cảnh

Người dùng	Sản phẩm	Đánh giá	Thời gian	Địa điểm	Bạn đồng hành
u_1	p_1	5	Cuối tuần	Tại nhà	Trẻ em
u_1	p_1	4	Trong tuần	Tại nhà	Gia đình
u_2	p_1	3	Cuối tuần	Tại rạp	Đối tác
u_2	p_1	4	Trong tuần	Tại nhà	Gia đình
u_3	p_1	3	Cuối tuần	Tại rạp	Đối tác
u_3	p_2	2	Cuối tuần	Tại rạp	Đối tác

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 [6][7]: 1) Lọc trước theo ngữ cảnh; 2) Lọc sau theo ngữ cảnh và 3) Mô hình hóa ngữ cảnh. Hình 2.3 minh họa trực quan cơ chế tích hợp ngữ cảnh vào hệ tư vấn theo 3 hướng này.



Hình 2.3. Các mô hình kết hợp ngữ cảnh vào hệ tư vấn [6]

2.3.3.1. Lọc trước theo ngữ cảnh

Hướng tiếp cận lọc trước theo ngữ cảnh (Contextual Prefiltering) thực hiện như sau: Từ tập dữ liệu đầu vào của hệ tư vấn dựa trên ngữ cảnh là ma trận đánh giá đa chiều, hệ thống sẽ sử dụng thông tin ngữ cảnh của người dùng hiện thời để lọc tập dữ liệu ban đầu nhằm chỉ giữ lại những dữ liệu phù hợp với ngữ cảnh yêu cầu [88]. Một số phương pháp lọc trước theo ngữ cảnh được đưa ra như: Phương pháp phân chia dữ liệu (Splitting) [89], phương pháp sử dụng một tập con các chiều ngữ cảnh để lọc (Context Relaxation) [90], phương pháp lọc dựa trên ngữ nghĩa (Semantic Filtering) [91] hoặc phương pháp phân cụm người dùng và sản phẩm dựa trên điều kiện ngữ cảnh [80].

Sau khi lọc thông tin ngữ cảnh, bài toán tư vấn theo ngữ cảnh dựa trên ma trận đánh giá đa chiều (theo công thức (2.8)) sẽ chuyển về bài toán tư vấn truyền thống dựa trên ma trận đánh giá hai chiều (theo công thức (2.7)). Tập dữ liệu lọc được sẽ dùng để huấn luyện và tư vấn. Một trong những lợi ích quan trọng nhất của hướng tiếp cận này là việc có thể sử dụng lại các phương pháp tư vấn truyền thống hai chiều đã được đề cập trong Mục 1.5 của Chương 1 để áp dụng cho hệ tư vấn theo ngữ cảnh. Tuy nhiên vấn đề đặt ra với hướng tiếp cận lọc trước theo ngữ cảnh cho hệ tư vấn là vấn đề dữ liệu thưa và ít khai thác được thông tin ngữ cảnh vào quá trình dự đoán và tư vấn.

Ví dụ trong hệ tư vấn phim, nếu một người dùng muốn xem một bộ phim vào thứ 7 thì chỉ những bộ phim được chiếu vào thứ 7 đã được xếp hạng mới được dùng để gợi ý cho người dùng.

2.3.3.2. Lọc sau theo ngữ cảnh

Trái ngược với hướng tiếp cận lọc trước ngữ cảnh, lọc sau theo ngữ cảnh (Contextual Postfiltering) sử dụng toàn bộ ma trận đánh giá đã loại bỏ đi các chiều ngữ cảnh để huấn luyện và tư vấn. Do vậy, các phương pháp tư vấn truyền thống sẽ được sử dụng để đưa ra tư vấn ngay từ đầu mà không phụ thuộc vào ngữ cảnh. Kết quả tư vấn sẽ được lọc lại một lần nữa theo ngữ cảnh của người dùng hiện thời để thu được kết quả tư vấn cuối cùng. Tuy nhiên, do sử dụng toàn bộ ma trận đánh giá đã loại bỏ đi các chiều ngữ cảnh để huấn luyện và tư vấn nên các thông tin ngữ cảnh cũng không được khai thác, đồng thời một số dữ liệu trùng lặp và nhiễu có thể ảnh hưởng tới chất lượng của các phương pháp tư vấn truyền thống.

Ví dụ trong hệ tư vấn phim, nếu một người dùng muốn xem một bộ phim vào cuối tuần. Sau khi thực hiện dự đoán danh sách phim gợi ý theo các phương pháp tư vấn truyền thống (bỏ qua yếu tố thông tin ngữ cảnh), hệ thống sẽ lọc loại bỏ các bộ phim không được chiếu vào cuối tuần từ danh sách phim gợi ý.

2.3.3.3. Mô hình hóa ngữ cảnh

Hướng tiếp cận mô hình hóa ngữ cảnh (Contextual Modeling) cho phép thông tin ngữ cảnh, người dùng và sản phẩm được biểu diễn trực tiếp trong cùng một mô hình. Khi đó ma trận đánh giá đa chiều sẽ được sử dụng trực tiếp cho quá trình huấn luyện và tư vấn. Với hướng tiếp cận này, một số thuật toán tư vấn theo ngữ cảnh được đưa ra có độ phức tạp lớn hơn các phương pháp tư vấn truyền thống. Các phương pháp mô hình hóa dựa vào ngữ cảnh được phân chia thành hai nhóm chính: Mô hình hóa ngữ cảnh độc lập và mô hình hóa ngữ cảnh phụ thuộc [88].

Phân rã Ten-xơ (Tensor Decomposition) [92] là một phương pháp điển hình thuộc nhóm phương pháp mô hình hóa ngữ cảnh độc lập. Phương pháp này cho phép biểu diễn người dùng, sản phẩm và các chiều ngữ cảnh trong một không gian đa chiều, mỗi chiều là độc lập nhau. Một vấn đề đối với phương pháp phân rã Ten-xơ là không gian lưu trữ cùng xử lý sẽ rất phức tạp khi số lượng chiều ngữ cảnh quá lớn, đồng thời mối liên quan giữa các chiều ngữ cảnh không được khai thác.

Không giống như phương pháp mô hình hóa ngữ cảnh độc lập coi ngữ cảnh không phụ thuộc vào người dùng và sản phẩm, phương pháp mô hình hóa ngữ cảnh phụ thuộc sẽ mô tả và khai thác sự phụ thuộc giữa người dùng, sản phẩm và ngữ cảnh tương ứng. Hai kỹ thuật được sử dụng trong phương pháp này đó là mô hình hóa dựa trên độ chênh lệch và mô hình hóa dựa trên độ tương tự [6]. Kỹ thuật mô hình hóa dựa trên độ chênh lệch sẽ thiết lập một mức chênh lệch cộng thêm vào đánh giá không có ngữ cảnh để suy ra đánh giá của người dùng cho sản phẩm trong một tình huống ngữ cảnh cụ thể. Kỹ thuật mô hình hóa dựa trên độ tương tự sẽ thiết lập một mức độ tương tự nhân với đánh giá không có ngữ cảnh để điều chỉnh đánh giá của người dùng cho sản phẩm trong một tình huống ngữ cảnh cụ thể. Theo đó, để đưa ra dự đoán đánh giá của người dùng với sản phẩm trong từng tình huống ngữ cảnh, có hai việc cần thực hiện đó là: (1) Lựa chọn phương pháp dự đoán đánh giá của người dùng với sản phẩm khi không có ngữ cảnh và (2) Lựa chọn độ đo tính mức chênh lệch hoặc mức tương tự giữa các tình huống ngữ cảnh. Với việc (1) về cơ bản chúng ta có thể sử dụng tất cả các phương pháp tư vấn truyền thống để thực hiện, trong đó hai phương pháp tư vấn cộng tác theo ngữ cảnh dựa trên Matrix

Factorization [77] và SLIM [93][94] được đánh giá là mang lại hiệu quả tương đối tốt. Với việc (2) chúng ta có thể sử dụng những độ đo khoảng cách Euclid, Minkowski…, độ đo tương tự Cosin, Entropy…, độ đo tương quan Pearson, Spearman, Kendal,… để tính toán mức độ chênh lệch hoặc tương tự giữa các tình huống ngữ cảnh. Thực nghiệm cho thấy các phương pháp mô hình hóa ngữ cảnh phụ thuộc cho kết quả tốt hơn phương pháp mô hình hóa ngữ cảnh độc lập trong nhiều trường hợp [95]. Tuy nhiên vấn đề đặt ra với các phương pháp mô hình hóa ngữ cảnh phụ thuộc khi tích hợp ngữ cảnh vào hệ tư vấn truyền thống là vấn đề dữ liệu thưa và khả năng mở rộng của nó.

Một hướng tiếp cận khác để tích hợp ngữ cảnh vào hệ tư vấn là dựa trên mô hình đồ thị. Nhiều nghiên cứu đã chỉ ra rằng mô hình đồ thị giải quyết khá tốt vấn đề dữ liệu thưa và khả năng mở rộng dữ liệu [96]. Neves ARM và các cộng sự của mình [97] đã đưa ra đề xuất phương pháp tư vấn theo ngữ cảnh trên cơ sở Ontology và kỹ thuật kích hoạt lan truyền (Spreading Activation). Punam Bedi và Richa [98] đề xuất một phương pháp tiếp cận mới nâng cao hiệu quả tư vấn cho người dùng trong các nhà hàng dựa vào Ontology và kỹ thuật kích hoạt lan truyền. Emrah và các cộng sự [99] đề xuất một hệ tư vấn theo ngữ cảnh dựa trên một mô hình đồ thị cộng tác cho các chương trình trên ti vi. Z. Bahramian và các cộng sự [100] đề xuất hệ tư vấn địa điểm du lịch theo ngữ cảnh dựa trên mô hình kích hoạt lan truyền. Tuy nhiên các nghiên cứu đã có này chủ yếu tập trung vào việc xây dựng các mô hình biểu diễn đồ thị áp dụng riêng lẻ cho từng bài toán tư vấn theo ngữ cảnh cụ thể, mà hầu như chưa có một giải pháp tổng thể chung cho các bài toán tư vấn theo ngữ cảnh.

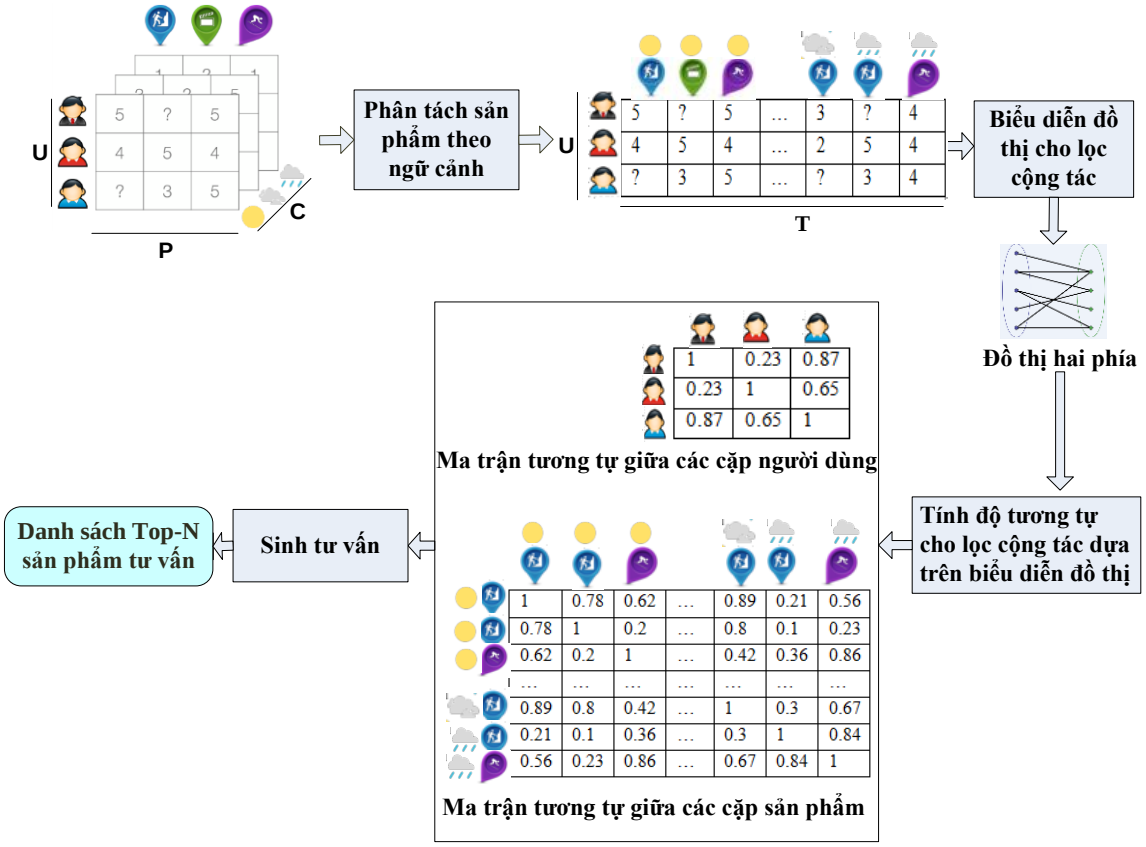
Qua phần trình bày ở trên về các phương pháp tích hợp ngữ cảnh vào hệ tư vấn đã chỉ ra rằng mỗi phương pháp đều có những ưu nhược điểm riêng, việc lựa chọn phương pháp nào sẽ phụ thuộc vào hiệu quả cho từng bộ dữ liệu của bài toán nghiệp vụ khác nhau. Căn cứ vào kết quả thực nghiệm nhiều nghiên cứu đã chỉ ra rằng không có phương pháp nào là tốt cho mọi trường hợp dữ liệu [76], nhưng hai hướng tiếp cận lọc trước theo ngữ cảnh và mô hình hóa ngữ cảnh đã và đang thu hút được sự quan tâm đặc biệt của cộng đồng nghiên cứu về hệ tư vấn theo ngữ cảnh, với số

lượng bài báo công bố lớn hơn hướng tiếp cận còn lại và chứng minh cho hiệu quả tư vấn cao trong nhiều trường hợp. Mặc dù vậy, một số vấn đề chính còn tồn tại với phương pháp thuộc hướng tiếp cận lọc trước ngữ cảnh và mô hình hóa ngữ cảnh là vấn đề dữ liệu thưa. Ngoài ra việc tích hợp các thông tin ngữ cảnh vào quá trình huấn luyện và tư vấn khiến cho các phương pháp mô hình hóa ngữ cảnh còn gặp phải vấn đề là tăng độ phức tạp tính toán khi số chiều dữ liệu tăng lên.

Để giảm thiểu những hạn chế nêu trên, luận án đề xuất một phương pháp tư vấn cộng tác theo ngữ cảnh mới thuộc hướng tiếp cận lọc trước ngữ cảnh, nhằm phát huy tính đơn giản trong cài đặt và tận dụng được các phương pháp tư vấn truyền thống đã có. Đồng thời phương pháp đề xuất cũng nhằm giải quyết hạn chế còn tồn tại phổ biến đối với hướng tiếp cận lọc trước ngữ cảnh, đó là vấn đề dữ liệu thưa và tích hợp hiệu quả thông tin ngữ cảnh vào quá trình tư vấn theo ngữ cảnh. Phương pháp đề xuất được trình bày chi tiết trong Mục 2.3.4 phát triển từ độ đo tương tự cho lọc cộng tác dựa trên mô hình đồ thị đề xuất trong Mục 2.2 của tác giả.

2.3.4. Phương pháp đề xuất

Nội dung phần này trình bày đề xuất về phương pháp tư vấn cộng tác mới cho hệ tư vấn theo ngữ cảnh được phát triển từ độ đo tương tự cho lọc cộng tác dựa trên mô hình đồ thị trình bày trong Mục 2.2. Về cơ bản phương pháp đề xuất được thực hiện bằng cách kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh; 2) Phương pháp k-láng giềng gần nhất với độ đo tương tự tính toán dựa trên mô hình đồ thị đề xuất (Mục 2.2). Sự kết hợp của hai phương pháp này trong phương pháp đề xuất được thể hiện qua bốn bước: 1) Phân tách sản phẩm theo ngữ cảnh; 2) Biểu diễn đồ thị cho lọc cộng tác; 3) Tính độ tương tự cho lọc cộng tác dựa trên biểu diễn đồ thị; 4) Sinh tư vấn. Bốn bước này kết hợp với nhau trong một bộ khung đề xuất về triển khai phương pháp lọc cộng tác dựa trên mô hình đồ thị cho hệ tư vấn theo ngữ cảnh trong Hình 2.5 dưới đây.



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

Việc thực hiện các bước được miêu tả chi tiết trong các phần dưới đây.

2.3.4.1. Phân tách sản phẩm theo ngữ cảnh

Thông tin đầu vào cho bài toán tư vấn theo ngữ cảnh được miêu tả trong 2.3.2 gồm có: Tập hợp hữu hạn gồm N người dùng $U = \{u_1, u_1, \dots, u_N\}$, M sản phẩm $P = \{p_1, p_2, \dots, p_M\}$ 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. Từ thông tin đầu vào trên, việc phân tách sản phẩm theo ngữ cảnh sẽ chuyển hóa sản phẩm ban đầu theo ngữ cảnh thành các sản phẩm giả lập. Mỗi sản phẩm giả lập được tạo ra từ sự kết hợp sản phẩm ban đầu với một tình huống ngữ cảnh cụ thể, thủ tục này gọi là "Item Splitting" [101].

Thủ tục "Item Splitting" trước đây chỉ sử dụng một chiều ngữ cảnh duy nhất để phân tách sản phẩm theo ngữ cảnh, việc chọn chiều ngữ cảnh này dựa vào độ đo thống kê, như độ lợi thông tin (Information gain) [77][102] . Điều này trong nhiều

trường hợp khiến cho khá nhiều thông tin từ các chiều ngữ cảnh khác không được khai thác triệt để vào quá trình tư vấn sau này.

Từ lập luận đưa ra ở trên, luận án đề xuất phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến nhằm khắc phục hạn chế nêu trên của phương pháp phân tách sản phẩm theo ngữ cảnh nguyên thủy. 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. Khi đó, mỗi điều kiện ngữ cảnh của C là sự kết hợp các điều kiện ngữ cảnh của các chiều tương ứng. Số lượng điều kiện ngữ cảnh của C là N_c, với $N_c = N_{c_1} * N_{c_2} * \dots * N_{c_K}$.
- **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. Khi đó, mỗi sản phẩm giả lập thuộc T là sự kết hợp của một sản phẩm ban đầu thuộc P với một điều kiện ngữ cảnh thuộc C. Số lượng sản phẩm trong tập T là H, với $H = M * N_c$.
- **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.

Ví dụ áp dụng phương pháp phân tách sản phẩm theo ngữ cảnh lên ma trận đánh giá đa chiều của lọc cộng tác theo ngữ cảnh (Bảng 2.4) ta thu được ma trận đánh giá hai chiều (Bảng 2.5), với t_1 là sản phẩm giả lập được tạo ra bởi sự kết hợp của sản phẩm p_1 và tình huống ngữ cảnh ("Cuối tuần","Tại nhà","Trẻ em"). Với ví dụ được đưa ra trong Bảng 2.4, hệ tư vấn có 2 sản phẩm và 12 tình huống ngữ cảnh có thể có, do vậy số lượng sản phẩm giả lập được sinh ra theo phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến là 24. Ma trận đánh giá hai chiều nhận được thể hiện trong Bảng 2.5, luận án sắp xếp những cặp người dùng - sản phẩm có đánh giá trong những dòng trên cùng của ma trận và những cặp còn lại không có đánh giá

ở bên dưới. Để tiết kiệm không gian trình bày, những cặp người dùng - sản phẩm không có đánh giá không nêu đầy đủ trong Bảng 2.5.

Bảng 2.5. Ma trận đánh giá hai chiều nhận được sau phân tách sản phẩm theo ngữ cảnh

Người dùng	Sản phẩm giả lập	Đánh giá
u_1	t_1	5
u_1	t_3	4
u_2	t_2	3
u_2	t_3	4
u_3	t_2	3
u_3	t_4	2
u_1	t_2	0
...	...	...
u_3	t_{24}	0

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). Trên thực tế, số lượng các đánh giá ban đầu đưa ra bởi người dùng cho các sản phẩm trong các tình huống ngữ cảnh là rất ít, khiến cho ma trận R_1 rất thưa. Khi áp dụng thủ tục phân tách sản phẩm theo ngữ cảnh cải tiến lên R_1, với việc giới thiệu các sản phẩm giả lập, sẽ càng khiến ma trận R_o thu được càng thưa thớt hơn nữa.

Ví dụ: Thông tin thu thập được cho hệ tư vấn theo ngữ cảnh thể hiện qua Bảng 2.4, có 6 đánh giá cho 2 sản phẩm cho trước. Sau quá trình phân tách sản phẩm theo ngữ cảnh được Bảng 2.5, có 6 đánh giá cho 24 sản phẩm giả lập. Như vậy, rất nhiều sản phẩm giả lập chưa được đánh giá, điều này khiến ma trận đánh giá 2 chiều nhận được sau phân tách ngữ cảnh càng thưa thớt hơn nữa.

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. Mô hình cho phép biểu diễn tất cả quan điểm của người dùng đối với các sản phẩm giả lập bằng một đồ thị. Khi đó, mức độ tương tự cho lọc cộng tác dựa trên mô hình đồ thị được tính toán trên cơ sở khai thác tất cả những mối quan hệ trực tiếp và bắc cầu giữa các cặp người dùng hoặc các cặp sản phẩm giả lập. Việc khai thác đầy đủ mối quan hệ bắc cầu giữa các đối tượng trong hệ thống trên đồ thị sẽ góp phần giải quyết vấn đề thưa dữ liệu và nâng cao hiệu quả dự đoán của phương pháp cộng tác cho hệ tư vấn theo ngữ cảnh. Các nội dung này lần lượt được đề cập đến trong Mục 2.3.4.2 và Mục 2.3.4.3 sau đây.

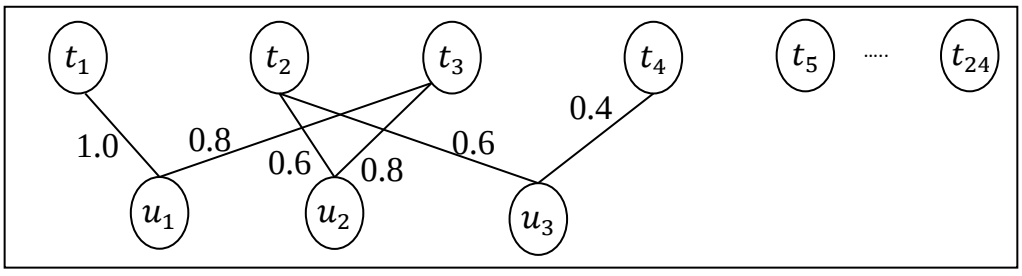
2.3.4.2. Biểu diễn đồ thị cho lọc cộng tác

Trước tiên, áp dụng công thức (2.1) được trình bày trong Mục 2.2.1 nhằm chuyển đổi dạng biểu diễn cho ma trận đánh giá hai chiều R_o từ Bảng 2.5 về Bảng 2.6 dưới đây (Các cột cuối không có đánh giá).

Bảng 2.6. Ma trận đánh giá chuyển đổi cho ma trận đánh giá 2 chiều của Bảng 2.5

	t_1	t_2	t_3	t_4	t_5	...	t_{24}
u_1	1	0	0.8	0	0	...	0
u_2	0	0.6	0.8	0	0	...	0
u_3	0	0.6	0	0.4	0	...	0

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 (Bảng 2.6). Khi đó, đồ thị hai phía biểu diễn cho lọc cộng tác được thể hiện trong Hình 2.5.



Hình 2.5. Đồ thị biểu diễn cho lọc cộng tác gồm tập người dùng và tập sản phẩm giả lập

Đồ thị hai phía nhận được không có các liên kết tới các đỉnh sản phẩm $t_5, \dots, t_{24}$ bởi vì các sản phẩm giả lập này không được đánh giá bởi bất kỳ người dùng nào.

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, đó là: 1) Tính toán mức độ tương tự giữa các cặp người dùng dựa trên đồ thị; 2) Tính toán mức độ tương tự giữa các cặp sản phẩm dựa trên đồ thị.

Việc làm này được xác định thông qua ma trận trọng số tổng quát biểu diễn đồ thị G dưới đây.

$$Z = \begin{pmatrix} UZ(N \times N) & W(N \times H) \\ W^T(H \times N) & TZ(H \times H) \end{pmatrix} \quad (2.9)$$

Với $Z = \{z_{ij}\}$ là ma trận trọng số biểu diễn đồ thị $G(i = 1,2, \dots (N + H);\ j = 1,2, \dots (N + H))$. Khi đó ma trận vuông Z sẽ được chia thành bốn phần theo công thức (2.9). Trong đó, $W(N \times H)$ được xác định theo công thức (2.3) biểu diễn mối quan hệ giữa người dùng và sản phẩm giả lập, $W^T(H \times N)$ là chuyển vị của $W(N \times H)$ biểu diễn mối quan hệ giữa sản phẩm giả lập và người dùng, ma trận vuông $UZ(N \times N)$ biểu diễn mối quan hệ giữa người dùng và người dùng, $TZ(H \times H)$ biểu diễn mối quan hệ giữa sản phẩm giả lập với sản phẩm giả lập. Các phần tử của ma trận $UZ(N \times N)$, $TZ(H \times H)$ ban đầu đều có giá trị 0, tương ứng với mức độ tương tự giữa các cặp người dùng hoặc giữa các cặp sản phẩm giả lập không xác định tại thời điểm ban đầu.

Sau khi xây dựng được ma trận Z việc tính mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm theo công thức (2.5), (2.6) đề xuất trong mục 2.2.2 của luận án.

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 ở trên trong 2 trường hợp dưới đây.

- **Trường hợp độ đo tương tự giữa các cặp người dùng được sử dụng bởi phương pháp User-Based k-NN để sinh tư vấn**

Với mỗi người dùng hiện thời u_a, gọi U_a là tập gồm K_1 người dùng tương tự nhất với u_a. Khi đó, đánh giá chưa biết của người dùng u_a với sản phẩm giả lập t_j (gọi là r_{aj}) được dự đoán căn cứ vào trung bình đánh giá của các người dùng trong tập U_a với sản phẩm t_j.

$$r_{aj} = \frac{\sum_{r_{ij} \in R_j} r_{ij}}{|R_j|},\ R_j = \{r_{ij} | r_{ij} \neq 0, u_i \in U_a\} \tag{2.10}$$

- **Trường hợp độ đo tương tự giữa các cặp sản phẩm được sử dụng bởi phương pháp Item-Based k-NN để sinh tư vấn**

Với mỗi sản phẩm giả lập t_j chưa được đánh giá bởi người dùng hiện thời u_a, gọi T_j là tập gồm K_1 sản phẩm tương tự nhất với t_j và đã được đánh giá bởi u_a. Khi đó đánh giá chưa biết của người dùng u_a với sản phẩm t_j được dự đoán căn cứ vào trung bình đánh giá của người dùng u_a với các sản phẩm thuộc tập T_j.

$$r_{aj} = \frac{\sum_{r_{ak} \in R_a} r_{ak}}{|R_a|},\ R_a = \{r_{ak} | r_{ak} \neq 0, t_k \in T_j\} \tag{2.11}$$

Từ dự đoán đánh giá r_{aj} được xác định theo một trong hai công thức (2.10) hoặc (2.11), hệ thống sẽ chọn ra K_2 các sản phẩm có dự đoán đánh giá cao để tư vấn cho người dùng hiện thời.

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 với 4 bước thực hiện trình bày ở trên, 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). Phần trình bày dưới đây sẽ trình bày đầy đủ hai thuật toán này.

Thuật toán lọc cộng tác theo ngữ cảnh dựa vào mức độ tương tự giữa các cặp người dùng trên mô hình đồ thị

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.*
- *$c \in (C_1 \times C_2 \times ... \times C_K)$ 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

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

$L \Leftarrow L + 2$; // 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

Thuật toán lọc cộng tác theo ngữ cảnh dựa vào mức độ tương tự giữa các cặp sản phẩm trên mô hình đồ thị

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.*
- *$c \in (C_1 \times C_2 \times ... \times C_K)$ 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

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

$L \Leftarrow L + 2$; // 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

Tại bước 2 của thuật toán IS-UserBased-Graph và IS-ItemBased-Graph, điều kiện dừng để tính được mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm được xác định căn cứ vào Định lý 2.1 và Định lý 2.2 nêu ra tại Mục 2.2.2.

2.4. Thực nghiệm và kết quả

Nội dung phần này trình bày đánh giá hiệu quả của phương pháp đề xuất trong sự so sánh với các phương pháp tư vấn theo ngữ cảnh phổ biến.

2.4.1. Dữ liệu thực nghiệm

Để thấy rõ hiệu quả của phương pháp đề xuất, tác giả thực hiện tiến hành thực nghiệm trên ba bộ dữ liệu DepaulMovie [103], MovieLens (https://grouplens.org/), InCarMusic [103].

- Bộ dữ liệu 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 là *Time, Location, Companion.* Chiều ngữ cảnh *Time* có 2 điều kiện ngữ cảnh ("Weekend", "Weekday"), chiều ngữ cảnh *Location* có 2 điều kiện ngữ cảnh

("Home", "Cinema"), chiều ngữ cảnh *Companion* có 3 điều kiện ngữ cảnh ("Alone", "Family", "Partner"). Các mức đánh giá nằm trong dải từ 1 đến 5, mức độ thưa thớt của dữ liệu là 94,516%. Các mức đánh giá 1, 2, 3, 4, 5 được chuyển đổi thành 0.2, 0.4, 0.6, 0.8, 1.0.

- Bộ dữ liệu 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 là *TimeOfDay, TimeOfWeek*. Chiều ngữ cảnh *TimeOfDay* có 5 điều kiện ngữ cảnh ("Morning", "Noon", "Afternoon", "Evening", "Night"), chiều ngữ cảnh *TimeOfWeek* có 2 điều kiện ngữ cảnh ("Weekend", "Weekday"). Các mức đánh giá nằm trong dải từ 1 đến 5, mức độ thưa thớt của dữ liệu là 93,89%. Các mức đánh giá 1, 2, 3, 4, 5 được chuyển đổi thành 0.2, 0.4, 0.6, 0.8, 1.0.
- Bộ dữ liệu 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 là *Driving style, Road type, Landscape, Sleepiness, Traffic conditions, Mood, Weather, Natural Phenomena.* Chiều ngữ cảnh *Driving style* có 2 điều kiện ngữ cảnh ("Relaxed driving", "Sport driving"), chiều ngữ cảnh *Road type* có 3 điều kiện ngữ cảnh ("City", "Highway", "Serpentine"), chiều ngữ cảnh *Landscape* có 4 điều kiện ngữ cảnh ("Coast line", "country side", "mountains/hills", "Urban"), chiều ngữ cảnh *Sleepiness* có 2 điều kiện ngữ cảnh ("Awake", "Sleepy"), chiều ngữ cảnh *Traffic conditions* có 3 điều kiện ngữ cảnh ("Free road", "Many Cars", "Traffic jam"), chiều ngữ cảnh *Mood* có 4 điều kiện ngữ cảnh ("Active", "Happy", "Lazy", "Sad"), chiều ngữ cảnh *Weather* có 4 điều kiện ngữ cảnh ("Cloudy", "Snowing", "Sunny", "Rainy"), chiều ngữ cảnh *Natural Phenomena* có 4 điều kiện ngữ cảnh ("Day time", "Morning", "Night", "Afternoon"). Các mức đánh giá nằm trong dải từ 1 đến 5, mức độ thưa thớt của dữ liệu là 99.9996996%. Các mức đánh giá 1, 2, 3, 4, 5 được chuyển đổi thành 0.2, 0.4, 0.6, 0.8, 1.0.

2.4.2. Cài đặt thực nghiệm

2.4.2.1. Độ đo

Hai nhiệm vụ chính của hệ tư vấn là dự đoán đánh giá và tư vấn danh sách ngắn các sản phẩm cho người dùng hiện thời. Để đánh giá hiệu quả của đánh giá dự đoán, các độ đo thường được sử dụng là $MAE, RMSE, MPE$. Để đánh giá hiệu quả tư vấn danh sách sản phẩm, các độ đo điển hình được sử dụng là $Precision@N$, $Recall@N$ và $MAP@N$.

Qua khảo sát, việc cài đặt các phương pháp tư vấn theo ngữ cảnh cơ sở hiện nay thường đưa ra trực tiếp danh sách các sản phẩm tư vấn mà không trả về kết quả dự đoán đánh giá, nên khi thực nghiệm đánh giá độ chính xác các thuật toán này dùng độ đo $Precision@N$, $MAP@N$ để đánh giá. Mặc dù để đưa ra tư vấn, phương pháp đề xuất bởi luận án tính qua 2 giai đoạn chính là dự đoán đánh giá và lấy kết quả đánh giá dự đoán làm cơ sở sinh ra tư vấn danh sách sản phẩm, tuy nhiên để thuận tiện trong việc so sánh với các phương pháp tư vấn theo ngữ cảnh cơ sở, tác giả sẽ sử dụng cùng độ đo đánh giá độ chính xác của danh sách tư vấn ($Precision@N$, $MAP@N$) để so sánh hiệu quả của phương pháp đề xuất với các phương pháp cơ sở. Chi tiết các độ đo $Precision@N$, $MAP@N$ được trình bày trong Mục 1.6.3 của luận án.

Trong thực nghiệm, luận án sử dụng $N = 10$. Giá trị $Precision@N$ và $MAP@N$ lớn thể hiện thuật toán tư vấn có độ chính xác càng cao.

2.4.2.2. Phương pháp thực nghiệm

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn, tác giả sử dụng lý thuyết về phương pháp đánh giá hệ thống được trình bày trong Mục 1.6.1 của luận án. Trong đó, cách phân chia tập dữ liệu U thành 2 tập U_{train} và U_{test} được sử dụng là phương pháp kiểm thử chéo (k-fold cross-validation) vì đây là phương pháp được sử dụng rộng rãi và cho kết quả đánh giá khách quan nhất. Trong thực nghiệm, luận án sẽ lấy $k = 10$ để tiến hành chia dữ liệu kiểm nghiệm. Việc thực nghiệm được thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.

2.4.2.3. Các phương pháp tư vấn được sử dụng để so sánh

- *BiasedMF* [29]: Phương pháp tư vấn không dựa vào ngữ cảnh, sử dụng mô hình BiasedMF để giải quyết bài toán lọc cộng tác.
- *UserSplitting-BiasedMF* [102]: Phương pháp tư vấn dựa vào ngữ cảnh, sử dụng phương pháp phân tách người dùng theo ngữ cảnh nguyên thủy, trong đó mỗi người dùng được tách thành hai người dùng giả lập tùy thuộc vào tình huống ngữ cảnh kết hợp với họ. Sau đó huấn luyện và đưa ra tư vấn sử dụng phương pháp phân rã ma trận BiasedMF (Biased-Matrix Factorization) [29].
- *ItemSplitting-BiasedMF* [101][104]: Phương pháp tư vấn dựa vào ngữ cảnh, sử dụng phương pháp phân tách sản phẩm theo ngữ cảnh nguyên thủy, trong đó mỗi sản phẩm được tách thành hai sản phẩm giả lập tùy thuộc vào tình huống ngữ cảnh kết hợp với nó. Sau đó huấn luyện và đưa ra tư vấn sử dụng phương pháp phân rã ma trận BiasedMF.
- *UISplitting-BasedMF* [104]: Phương pháp tư vấn dựa vào ngữ cảnh, sử dụng phương pháp phân tách cả người dùng và sản phẩm theo ngữ cảnh, sau đó huấn luyện và đưa ra tư vấn sử dụng phương pháp phân rã ma trận BiasedMF.
- *SLIM* [105] (Sparse Linear Method): Phương pháp tư vấn không dựa vào ngữ cảnh, sử dụng mô hình *SLIM* để giải quyết bài toán lọc cộng tác cho hệ tư vấn.
- *CSLIM* [93]: Phương pháp tư vấn dựa vào ngữ cảnh, mở rộng từ phương pháp tuyến tính thưa SLIM [105] cho hệ tư vấn theo ngữ cảnh. Trong thực nghiệm luận án lựa chọn phương pháp SLIM kết hợp với mô hình hóa ngữ cảnh dựa trên độ tương quan đa chiều MCS (Multidimensional Context Similarity) – gọi tắt là CSLIM_MCS, biến thể của CSLIM, được đánh giá cho hiệu quả tư vấn cao trong nhiều trường hợp để so sánh đánh giá.
- *ItemSplitting-SLIM* [106]: Phương pháp tư vấn dựa vào ngữ cảnh, sử dụng phương pháp phân tách sản phẩm theo ngữ cảnh nguyên thủy, trong đó mỗi sản phẩm được tách thành hai sản phẩm giả lập tùy thuộc vào tình huống ngữ cảnh kết hợp với nó. Sau đó huấn luyện và đưa ra tư vấn sử dụng phương pháp tuyến tính thưa SLIM.

- *UserBased-Graph*: Phương pháp tư vấn không dựa vào ngữ cảnh, sử dụng: kết hợp phương pháp tính độ tương tự giữa các cặp người dùng dựa trên mô hình đồ thị (Theo đề xuất mục 2.2.2.1 của luận án) với phương pháp User-Based k-NN để thực hiện huấn luyện và đưa ra tư vấn.
- *ItemBased-Graph:* Phương pháp tư vấn không dựa vào ngữ cảnh, sử dụng: kết hợp phương pháp tính độ tương tự giữa các cặp sản phẩm dựa trên mô hình đồ thị (Theo đề xuất mục 2.2.2.2 của luận án) với phương pháp Item-Based k-NN để thực hiện huấn luyện và đưa ra tư vấn.
- *ItemSplitting-UserBased-Graph*: Phương pháp lọc cộng tác theo ngữ cảnh kết hợp hai phương pháp: 1) *ItemSplitting* - Phân tách sản phẩm theo ngữ cảnh nguyên thủy (Dùng 1 chiều ngữ cảnh); 2) *UserBased-Graph*.
- *ItemSplitting-ItemBased-Graph:* Phương pháp lọc cộng tác theo ngữ cảnh kết hợp hai phương pháp: 1) *ItemSplitting* - Phân tách sản phẩm theo ngữ cảnh nguyên thủy (Dùng 1 chiều ngữ cảnh); 2) *ItemBased-Graph.*
- *IS-UserBased-Graph*: Phương pháp lọc cộng tác theo ngữ cảnh kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh cải tiến; 2) *UserBased-Graph*. Đây là phương pháp luận án đề xuất trình bày trong Mục 2.3.4.
- *IS-ItemBased-Graph*: Phương pháp lọc cộng tác theo ngữ cảnh kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh cải tiến; 2) *ItemBased-Graph*. Đây là phương pháp luận án đề xuất trình bày trong Mục 2.3.4.
- *IS-Graph*: (Phương pháp tư vấn dựa vào ngữ cảnh) sử dụng phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến, sau đó huấn luyện và đưa ra tư vấn sử dụng phương pháp lọc cộng tác dựa trên mô hình đồ thị được đề xuất bởi Huang và các cộng sự [35]. Trong thực nghiệm độ dài đường đi từ đỉnh người dùng tới đỉnh sản phẩm trên đồ thị được giới hạn là 5 để cho kết quả tốt nhất.

Để thực nghiệm các phương pháp này, tác giả sử dụng máy tính có cấu hình: Intel Core i7-3770 CPU và 8GB RAM. Các thuật toán tư vấn theo ngữ cảnh cơ sở đã có được thực nghiệm dựa vào bộ thư viện CARSKIT [103], các thuật toán tư vấn

đề xuất do tác giả tự cài đặt và thực nghiệm theo cùng chiến lược thực nghiệm với phương pháp cơ sở, đó là kiểm thử chéo (k-fold cross-validation, $k = 10$) và dùng cùng các độ đo đánh giá độ chính xác danh sách dự đoán phổ biến (Precision@10, MAP@10) để so sánh. Để đảm bảo đánh giá công bằng phương pháp đề xuất so với các phương pháp cơ sở, tác giả thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.

Các phương pháp tư vấn theo ngữ cảnh cơ sở được thực nghiệm với các giá trị tham số khác nhau và lựa chọn kết quả độ chính xác tốt nhất để so sánh với các phương pháp đề xuất bởi luận án. Chi tiết về miền giá trị cho các tham số của các phương pháp tư vấn theo ngữ cảnh cơ sở trình bày trong những công trình nghiên cứu liên quan được tham chiếu ở trên.

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	0.121	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 2.8. 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 2.9. 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 đưa ra trong Bảng 2.7, Bảng 2.8, Bảng 2.9 như sau:

1) Các phương pháp lọc cộng tác cho hệ tư vấn không sử dụng ngữ cảnh: phương pháp lọc cộng tác sử dụng độ đo tương tự dựa trên mô hình đồ thị đề xuất bởi luận án (Mục 2.2), đó là *UserBased-Graph*, *ItemBased-Graph* cho Precision@10 và MAP@10 cao hơn *BiasedMF* và *SLIM* trên cả 3 tập dữ liệu. Điều đó khẳng định 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 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 khá tốt. Mặc dù, phương pháp *UserSplitting-BiasedMF* cho giá trị $MAP@10$ lớn nhất so với các phương pháp còn lại trên tập dữ liệu DepaulMovie nhưng không thể khẳng định một trong ba phương pháp *UserSplitting-BiasedMF*, *ItemSplitting-BiasedMF*, *UISplitting-BiasedMF* là cho chất lượng tư vấn tốt hơn các phương pháp còn lại trong mọi bộ dữ liệu. Đồng thời, kết hợp phương pháp phân tách theo ngữ cảnh với phương pháp *BiasedMF* cho lại độ chính xác tốt hơn phương pháp *BiasedMF* thuần túy cho lọc cộng tác. Điều này hoàn toàn phù hợp với những nghiên cứu trước đây [102].
3) Kết quả thực nghiệm cũng chỉ ra rằng 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* ở 2 trong 3 tập dữ liệu. Đồng thời khi so sánh *ItemSplitting-UserBased-Graph, ItemSplitting-ItemBased-Graph* với các phương pháp tư vấn theo ngữ cảnh cơ sở cùng hướng sử dụng kết hợp *ItemSplitting* cũng cho kết quả tương tự, tức là Precision@10 có giá trị tốt hơn nhưng MAP@10 cho giá trị thấp hơn. 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) Kết hợp phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến (sử dụng đồng thời nhiều chiều ngữ cảnh) và phương pháp *UserBased-Graph, ItemBased-Graph* để tạo thành phương pháp đề xuất ***IS-UserBased-Graph***, ***IS-ItemBased-Graph***. Trong đó, căn cứ trên thực nghiệm các phương pháp đề xuất ở trên, tác giả chọn giá trị $L = 6$ và số lượng phần tử trong tập láng giềng là 30 để cho kết quả tốt nhất. So sánh về giá trị Precision@10 nhận thấy phương pháp đề xuất *IS-UserBased-Graph, IS-ItemBased-Graph* cho Precision@10 tương đương với *ItemSplitting-UserBased-Graph, ItemSplitting-ItemBased-Graph* ở 2 trong 3 tập dữ liệu (DepaulMovie và InCarMusic) và cao hơn hẳn đối với tập dữ liệu MovieLens. So sánh về giá trị MAP@10 của phương pháp đề xuất lớn hơn *ItemSplitting-UserBased-Graph, ItemSplitting-ItemBased-Graph* trong cả 3 tập dữ liệu. Điều đó chứng tỏ 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) Quan sát kết quả thực nghiệm của việc kết hợp giữa phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến (sử dụng đồng thời nhiều chiều ngữ cảnh) với phương pháp lọc cộng tác dựa trên mô hình đồ thị được đề xuất bởi Huang và các cộng sự [35] (*IS-Graph*) cho hệ tư vấn theo ngữ cảnh thấy rằng hiệu năng của phương pháp *IS-Graph* thay đổi tùy thuộc vào bộ dữ liệu. So sánh hiệu năng của phương pháp này với các phương pháp tư vấn theo ngữ cảnh cơ sở, chúng ta nhận thấy phương pháp IS-Graph cho giá trị $Precision@10$ cao hơn hẳn các phương pháp tư vấn theo ngữ cảnh dựa trên giải thuật MF và SLIM đã đề cập ở trên khi xét trên tập dữ liệu DepaulMovie và MovieLens, nhưng lại thấp hơn đối với tập dữ liệu InCarMusic. Sự không ổn định về mặt hiệu năng của phương pháp IS-Graph trong sự so sánh với các phương pháp khác khi xét với các bộ dữ liệu khác nhau cho thấy việc tính mức độ phù hợp của sản phẩm chưa được đánh giá với người dùng có lợi điểm là cho phép khai thác mối quan hệ bắc cầu giữa đỉnh người dùng và đỉnh sản phẩm dựa trên đồ thị, giúp hạn chế ảnh hưởng vấn đề dữ liệu thưa. Tuy nhiên từ sự biến động của kết quả kiểm nghiệm đối với phương pháp IS-Graph, có thể khẳng định rằng cách tiếp cận của phương pháp này có thể không phải là cách tốt nhất để khai thác mối quan hệ bắc cầu giữa các

đỉnh trên đồ thị cho mục đích tư vấn. 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.

Để đưa ra mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm cho lọc cộng tác, luận án đề xuất biểu diễn đồ thị cho ma trận đánh giá hai chiều người dùng – sản phẩm và xây dựng một độ đo tương tự mới giữa các cặp người dùng hoặc các cặp sản phẩm dựa trên mô hình biểu diễn đồ thị này. Việc xác định mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm khi đó sẽ dựa trên các mối quan hệ trực tiếp và bắc cầu giữa các đỉnh người dùng hoặc giữa các đỉnh sản phẩm trên đồ thị, điều này giúp hạn chế ảnh hưởng của vấn đề dữ liệu thưa của lọc cộng tác. Đây chính là ưu điểm của độ đo tương tự đề xuất so với các độ đo tương tự dựa vào bộ nhớ trước đây vào việc giải quyết bài toán lọc cộng tác theo bộ nhớ cho hệ tư vấn truyền thống.

Trên cơ sở độ đo tương tự giữa các cặp người dùng hoặc giữa các cặp sản phẩm dựa trên mô hình đồ thị, luận án đưa ra đề xuất phương pháp tư vấn theo ngữ cảnh mới với hai thuật toán IS-UserBased-Graph, IS-ItemBased-Graph. Phương pháp tư vấn theo ngữ cảnh đề xuất cho phép khai thác đầy đủ thông tin ngữ cảnh bằng việc sử dụng thủ tục phân tách sản phẩm cải tiến theo ngữ cảnh nhằm chuyển hóa sản phẩm ban đầu thành các sản phẩm giả lập. Các sản phẩm giả lập thu nhận được lệ thuộc vào sự kết cặp của các sản phẩm ban đầu trong từng tình huống ngữ cảnh cụ thể, sự đa dạng của các tình huống ngữ cảnh kéo theo tập sản phẩm giả lập thu được

có kích thước lớn hơn khá nhiều so với tập sản phẩm ban đầu. Việc này cũng đồng thời giúp chuyển hóa ma trận đánh giá đa chiều ban đầu ($U \times P \times C \rightarrow R$) về dạng ma trận đánh giá hai chiều ($U \times T \rightarrow R$). Ma trận đánh giá hai chiều thu được giúp đơn giản trong cài đặt và cho phép tái sử dụng các phương pháp lọc cộng tác của hệ tư vấn truyền thống để sinh tư vấn. Trong đề xuất của mình, tác giả không sử dụng những phương pháp lọc cộng tác cho hệ tư vấn truyền thống đã biết lên ma trận đánh giá hai chiều giả lập để suy ra dự đoán cho hệ tư vấn theo ngữ cảnh. Ở đây, tác giả đã khai thác chính nghiên cứu của mình về độ đo tương tự giữa các cặp người dùng hoặc các cặp sản phẩm cho hệ tư vấn cộng tác truyền thống đề xuất trong Mục 2.2 của luận án, kết hợp với phương pháp kNN vào ma trận đánh giá hai chiều giả lập để thực hiện huấn luyện và sinh dự đoán đánh giá của người dùng cho các sản phẩm giả lập. Từ danh sách sản phẩm giả lập, ánh xạ ngược lại ta thu được các sản phẩm gốc kèm tình huống ngữ cảnh tương ứng. Sau đó, chọn những sản phẩm có đánh giá cao bởi người dùng hiện thời trong tình huống ngữ cảnh cho trước để đưa ra tư vấn theo ngữ cảnh cho người dùng.

Như vậy 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

Nội dung chương 3 trình bày kết quả nghiên cứu của luận án về đề xuất phương pháp hạn chế ảnh hưởng của vấn đề dữ liệu thưa cho lọc cộng tác bằng đồng huấn luyện (Co-Training), làm cơ sở đề xuất phương pháp lọc kết hợp bằng đồng huấn luyện. Phương pháp lọc kết hợp đề xuất phát triển từ phương pháp lọc cộng tác bằng đồng huấn luyện cho phép giải quyết vấn đề dữ liệu thưa, đồng thời tích hợp đầy đủ thông tin người dùng, sản phẩm và đánh giá của người dùng với sản phẩm vào quá trình dự đoán đánh giá, từ đó nâng cao chất lượng hệ tư vấn.

Trước hết, các hạn chế của việc giải quyết vấn đề dữ liệu thưa và tích hợp hiệu quả các thông tin liên quan trong các phương pháp lọc kết hợp đã có được trình bày trong Mục 3.1. Mục này cũng đồng thời nêu ra hướng đề xuất phương pháp lọc kết hợp mới của luận án trong sự nhìn nhận với những nghiên cứu liên quan. Tiếp theo, mục 3.2 trình bày đề xuất phương pháp lọc cộng tác bằng đồng huấn luyện. Trên cơ sở đề xuất đó, luận án đề xuất phương pháp lọc kết hợp bằng đồng huấn luyện trong Mục 3.3. Mục 3.4 trình bày về kết quả thực nghiệm, so sánh và đánh giá phương pháp đề xuất trong sự so sánh với các phương pháp tư vấn cơ sở. Mục cuối cùng 3.5 là kết luận và hướng nghiên cứu tiếp theo.

Nội dung Mục 3.2 trình bày trong chương được tổng hợp từ kết quả nghiên cứu [C2]; Nội dung Mục 3.3 được tổng hợp từ kết quả nghiên cứu [C5][C6][J1].

3.1. Đặt vấn đề

Như đã trình bày trong Chương 1, 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. Có bốn hướng tiếp cận chính để giải quyết bài toán lọc kết hợp cho hệ tư vấn truyền thống được đề cập trong Mục 1.5.3 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 cho cả lọc cộng tác và lọc nội dung. Trong các hướng tiếp cận trên, luận án tiếp cận hướng 2 về 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. Lý do cho hướng tiếp cận này của luận án là việc dịch chuyển bài toán lọc kết hợp về bài toán lọc cộng tác dựa vào bộ nhớ sẽ phát huy tối đa ưu điểm của phương pháp lọc cộng tác dựa vào bộ nhớ trong các hệ tư vấn thực tế [14][46][53]. Đồng thời, tiếp cận lọc kết hợp theo hướng này còn giúp tích hợp đầy đủ thông tin đánh giá của người dùng với sản phẩm, nội dung người dùng và nội dung sản phẩm vào quá trình tư vấn, từ đó cải thiện chất lượng tư vấn.

Tuy nhiên khi chuyển dịch bài toán lọc kết hợp về bài toán lọc cộng tác dựa vào bộ nhớ theo hướng tiếp cận kết hợp đặc tính của lọc nội dung vào lọc cộng tác thì tất cả các thông tin đầu vào của hệ tư vấn là một ma trận đánh giá duy nhất. Ma trận đánh giá hợp nhất này thể hiện đầy đủ mối quan hệ giữa người dùng, sản phẩm, đặc trưng người dùng, đặc trưng sản phẩm. Như vậy vấn đề mấu chốt trong các phương pháp lọc kết hợp theo hướng này là: (1) Lựa chọn phương pháp trích chọn đặc trưng phù hợp, sau đó là (2) Giải quyết vấn đề dữ liệu thưa trên ma trận đánh giá hợp nhất. Để giải quyết bước (1), các nghiên cứu đã có thường lựa chọn phương pháp TF-IDF trong việc xây dựng hồ sơ người dùng, khi đó véc tơ đặc trưng xây dựng được cho mỗi người dùng hoặc sản phẩm là một số thực trải đều trong khoảng [0,1] có miền giá trị không đồng nhất với miền giá trị đánh giá người dùng – sản phẩm, điều này khiến các đặc trưng nội dung chưa được kết hợp hiệu quả vào lọc cộng tác. Ngoài vấn đề thiếu sự kết hợp hiệu quả các đặc trưng nội dung vào lọc cộng tác, các nghiên cứu đã có theo hướng này vẫn đối mặt với vấn đề dữ liệu thưa của ma trận đánh giá hợp nhất cần phải tiếp tục nghiên cứu.

Nhằm phát huy thế mạnh của lọc kết hợp theo hướng tiếp cận kết hợp đặc tính của lọc nội dung vào lọc cộng tác dựa vào bộ nhớ, đồng thời khắc phục nhược điểm về ảnh hưởng vấn đề dữ liệu thưa và tích hợp hiệu quả các đặc trưng nội dung vào lọc cộng tác trong các phương pháp cùng hướng đã có, trong Mục 3.2 tiếp theo luận án trình bày đề xuất một phương pháp mới giải quyết vấn đề dữ liệu thưa cho lọc

cộng tác bằng đồng huấn luyện. Đồng huấn luyện thực hiện đồng thời hai mô hình phân lớp độc lập khi quan sát dữ liệu theo người dùng và theo sản phẩm nhằm đưa ra những đánh giá dự đoán đóng vai trò chia sẻ và bổ sung thông tin giữa hai mô hình phân lớp khác nhau, góp phần nâng cao kết quả dự đoán và hạn chế ảnh hưởng của vấn đề dữ liệu thưa trong lọc cộng tác. 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 [25][26][68][107]. Dựa trên đánh giá của người dùng với các sản phẩm của hệ thống, một mô hình phân lớp sẽ được xây dựng và huấn luyện để đưa ra dự đoán đánh giá của người dùng hiện thời với một sản phẩm cụ thể. Do vậy, việc xác định được phương pháp phân lớp phù hợp cho lọc cộng tác sẽ quyết định chất lượng của hệ tư vấn. Bài toán lọc cộng tác bằng phân lớp được phát biểu như sau.

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

Cho ma trận đánh giá $R = [r_{ix}]$ với $i = 1, 2, \dots, N$; $x = 1, 2, \dots, M$ thể hiện mối quan hệ giữa tập người dùng U và tập sản phẩm P như được trình bày ở Mục 1.3. Các hàng của ma trận tương ứng với tập người dùng, các cột của ma trận tương ứng với tập sản phẩm, các phần tử r_{ix} của ma trận tương ứng với đánh giá của người dùng u_i đối với sản phẩm p_x. Thông thường, mỗi người dùng chỉ đánh giá một tập rất nhỏ các sản phẩm, do vậy đa số các giá trị $r_{ix} = 0$. Nhiệm vụ của lọc cộng tác là điền vào hay dự đoán các giá trị thích hợp cho các giá trị này của 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 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

Về cơ bản, bài toán phân lớp là một loại bài toán của lĩnh vực học máy. Các nghiên cứu về học máy dựa trên phương thức học dữ liệu chỉ ra rằng có bốn hướng tiếp cận học máy chính [16][108], đó là: 1) Học có giám sát (Supervised learning); 2) Học không giám sát (Unsupervised learning); 3) Học bán giám sát (Semi-supervised learning); 4) Học củng cố (Reinforcement learning). Trong đó, học có giám sát là phương pháp học từ tập dữ liệu huấn luyện ban đầu hoàn toàn được gán nhãn từ trước để đưa ra dự đoán nhãn cho các mẫu dữ liệu mới, do vậy học có giám sát thường được sử dụng trong các bài toán phân lớp, phân loại (Classification). Trái với học có giám sát, học không giám sát là phương pháp học từ tập dữ liệu huấn luyện ban đầu hoàn toàn chưa được gán nhãn, phương pháp này thường được sử dụng cho lớp bài toán phân cụm (Clustering). Để kết hợp cả dữ liệu có nhãn và chưa có nhãn trong tập dữ liệu huấn luyện ban đầu, các phương pháp học bán giám sát được đưa ra. Tuỳ vào từng mục đích cụ thể, học bán giám sát có thể được áp dụng cho bài toán phân lớp hoặc phân cụm. Hướng tiếp cận học máy cuối cùng được nói đến ở đây là học củng cố, theo cách tiếp cận này việc học từ dữ liệu được thực hiện lặp đi lặp lại, tại mỗi lần lặp các thông tin phản hồi từ môi trường bên ngoài được đưa vào kết hợp với dữ liệu gốc để thực hiện huấn luyện dữ liệu đã biết. Trong bốn hướng tiếp cận học máy chính nêu ra ở đây thì phương pháp học có giám sát và bán giám sát là hai hướng tiếp cận phù hợp để giải quyết bài toán phân lớp ở quy mô tổng quát.

Xét mức độ phù hợp của hai hướng tiếp cận học máy (học có giám sát, học bán giám sát) đề cập ở trên 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, nếu áp dụng các phương pháp học máy có giám sát thì chỉ có một số ít

đánh giá tham gia vào quá trình học để sinh ra tư vấn, các giá trị đánh giá biết trước này còn gọi là nhãn phân loại. Như vậy, việc áp dụng các phương pháp học có giám sát cho hệ tư vấn cộng tác dựa vào bộ nhớ sẽ bỏ qua rất nhiều các mẫu dữ liệu khác chưa được gán nhãn vào quá trình tư vấn, vấn đề dữ liệu thưa này sẽ ảnh hưởng trực tiếp tới chất lượng tư vấn. Với mong muốn 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.

Có rất nhiều phương pháp học bán giám sát đã được đưa ra [109][110]. Trong đó các phương pháp điển hình thường được sử dụng như: phương pháp tự huấn luyện Selft-Training, phương pháp đồng huấn luyện Co-Training, phương pháp học đa khung nhìn (Multi-view learning) và các phương pháp học bán giám sát dựa trên đồ thị. Không có câu trả lời chính xác cho câu hỏi phương pháp nào là tốt nhất, điều này phụ thuộc vào sự phù hợp của tập dữ liệu đầu vào với mô hình học được lựa chọn.

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. Nội dung cụ thể của phương pháp được trình bày trong Mục 3.2.2.3.

3.2.2.3. Bán giám sát bằng phương pháp đồng huấn huyện

Đồng huấn luyện (Co-Training) [111][112][113] là phương pháp học bán giám sát điển hình mà các nhà khoa học đầu tư nghiên cứu. Phương pháp đồng huấn luyện dựa trên giả thuyết rằng các đặc trưng của tập dữ liệu huấn luyện có thể được phân chia thành 2 tập đặc trưng con [113]. Trường hợp lý tưởng là hai tập con này thoả mãn điều kiện độc lập nhau. Mỗi tập con phù hợp để huấn luyện một bộ phân lớp tốt. Thủ tục học được tiến hành như sau:

- Dùng 2 bộ phân lớp phù hợp để học các mẫu dữ liệu được biểu diễn bởi 2 tập đặc trưng con tương ứng.
- Mỗi bộ phân lớp thực hiện gán nhãn cho các mẫu dữ liệu chưa có nhãn, thu được kết quả chính là tập các mẫu dữ liệu đó kèm theo nhãn dự đoán của chúng. Trong tập kết quả của bộ phân lớp 1, chọn ra những mẫu dữ liệu (kèm nhãn đã dự đoán) có độ tin cậy cao nhất bổ sung vào tập huấn luyện của bộ phân lớp 2 và ngược lại.
- Mỗi bộ phân lớp được học lại tập dữ liệu huấn luyện (gồm dữ liệu gán nhãn ban đầu và dữ liệu gán nhãn mới bổ sung từ kết quả của bộ phân lớp kia). Quá trình được lặp lại cho đến khi tập dữ liệu chưa gán nhãn rỗng hoặc số vòng lặp đạt tới một ngưỡng xác định trước.

Phương pháp đồng huấn luyện thể hiện dưới dạng thuật toán như sau:

Đầu vào:

- D^L là tập các mẫu dữ liệu đã được gán nhãn, $D^L = \{x_i, y_i\}_{i=1}^{L}$
- D^U là tập các mẫu dữ liệu chưa được gán nhãn, $D^U = \{x_j\}_{j=L+1}^{L+U}$
- Mỗi mẫu dữ liệu được quan sát dưới hai góc nhìn $x = [x^{(1)}\ x^{(2)}]$.
- K là số lượng mẫu được gán nhãn cho mỗi lần lặp.
- Lựa chọn hai mô hình phân lớp $f^{(1)}, f^{(2)}$ để học các mẫu dữ liệu độc lập từ 2 quan sát $x^{(1)},\ x^{(2)}$.

Đầu ra:

- Tập D^U gồm các mẫu dữ liệu được gán nhãn.

Các bước thực hiện:

1. Từ D^L tạo 2 tập các mẫu dữ liệu đã được gán nhãn D^{L_1} và D^{L_2} theo 2 quan sát $x^{(1)}$ và $x^{(2)}$. Trong đó, $D^{L_1} = \{x_i^{(1)}, y_i\}_i^L$ và $D^{L_2} = \{x_i^{(2)}, y_i\}_i^L$.
2. Quan sát theo $x^{(1)}$, sử dụng tập dữ liệu D^{L_1} để huấn luyện bộ phân lớp $f^{(1)}$. Lựa chọn K mẫu $(x, f^{(1)}(x))$ có độ tin cậy cao nhất được gán nhãn bởi $f^{(1)}$ tới tập dữ liệu D^{L_2}, đồng thời loại bỏ K mẫu dữ liệu này khỏi D^U.
3. Quan sát theo $x^{(2)}$, sử dụng tập dữ liệu D^{L_2} để huấn luyện bộ phân lớp $f^{(2)}$. Lựa chọn K mẫu $(x, f^{(2)}(x))$ có độ tin cậy cao nhất được gán nhãn bởi $f^{(2)}$ tới tập dữ liệu D^{L_1}, đồng thời loại bỏ K mẫu dữ liệu này khỏi D^U.
4. Lặp lại các quá trình trên tới khi gán nhãn được hết các mẫu dữ liệu chưa có nhãn trong D^U hoặc số vòng lặp đạt đến ngưỡng xác định trước.

Thuật toán 3.1. Thuật toán đồng huấn luyện Co-Training

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. Với các giả định được nêu ra ở trên, Blum và Mitchell [111] chứng minh rằng phương pháp đồng huấn luyện là phương pháp học bán giám sát cho phép gán nhãn các mẫu dữ liệu chưa có nhãn từ một tập ít các mẫu dữ liệu có nhãn ban đầu với độ chính xác cao. Đây chính là sự khác biệt của phương pháp đồng huấn luyện so với các phương pháp học truyền thống khi thường bỏ qua sự phân chia và gộp chung tất cả các đặc trưng dữ liệu với nhau.

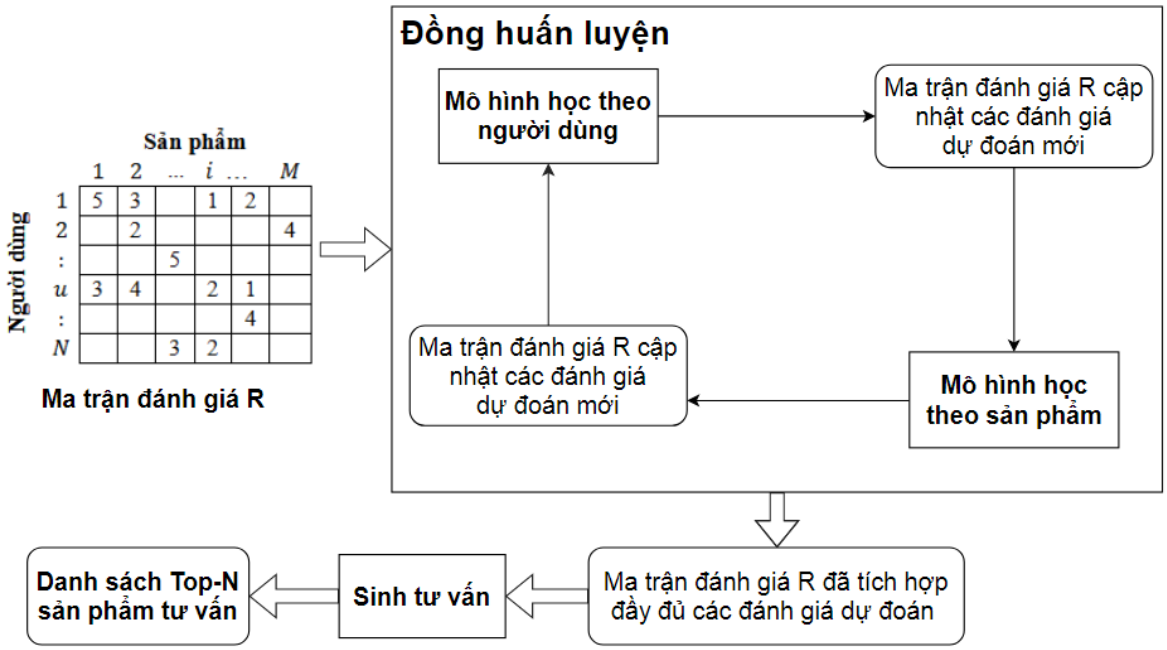
3.2.3. Mô hình đồng huấn luyện cho lọc cộng tác

Việc hiện thực hóa phương pháp đồng huấn luyện để giải quyết bài toán lọc cộng tác sẽ quan sát mỗi mẫu dữ liệu theo 2 cơ chế độc lập nhau $x = [x^{(1)}\ x^{(2)}]$, với $x^{(1)}$ là quan sát theo người dùng và $x^{(2)}$ là quan sát theo sản phẩm. Các nhãn y_i

tương ứng của các mẫu dữ liệu x_i chính là các giá trị đánh giá tương ứng của người dùng ($x_i^{(1)}$) với sản phẩm ($x_i^{(2)}$) từ ma trận R của lọc cộng tác. Như vậy tập hợp các mẫu dữ liệu có nhãn D^L và không có nhãn D^U được hợp nhất biểu diễn đầy đủ thông qua ma trận đánh giá R duy nhất, đây chính là thông tin đầu vào cho thuật toán đồng huấn luyện.

Quá trình đồng huấn luyện sẽ sử dụng 2 bộ phân lớp xác định $f^{(1)}, f^{(2)}$ nhằm học các mẫu dữ liệu độc lập từ quan sát theo người dùng ($x^{(1)}$) và quan sát theo sản phẩm ($x^{(2)}$) để gán nhãn cho các mẫu dữ liệu chưa biết trong tập D^U, trong trường hợp này là đưa ra dự đoán đánh giá cho những giá trị đánh giá chưa biết trong ma trận đánh giá R. Quá trình học này được lặp lại luân phiên giữa hai cơ chế quan sát theo người dùng và theo sản phẩm đến khi thỏa mãn điều kiện các mẫu dữ liệu đều được gán nhãn (Ma trận đánh giá R được cập nhật đầy đủ các giá trị đánh giá) hoặc số vòng lặp đạt đến ngưỡng xác định. Cụ thể quá trình học theo người dùng sẽ dự đoán được một số nhãn phân loại tin cậy cho mẫu dữ liệu chưa biết đánh giá chuyển giao cho quá trình học theo sản phẩm. Ngược lại, quá trình học theo sản phẩm cũng dự đoán được một số nhãn phân loại cho mẫu dữ liệu chưa biết đánh giá chuyển giao cho quá trình học theo người dùng. Mỗi quá trình học đó sẽ cập nhật những đánh giá dự đoán mới vào ma trận R, cũng đồng nghĩa với việc loại bỏ các mẫu dữ liệu vừa được gán nhãn khỏi tập hợp các mẫu dữ liệu không có nhãn D^U. Hai quá trình học được thực hiện luân phiên nhau theo đúng tinh thần của thuật toán đồng huấn luyện, điều này góp phần hạn chế ảnh hưởng của vấn đề dữ liệu thưa cho lọc cộng tác.

Hình 3.1 mô tả bộ khung đề xuất về triển khai lọc cộng tác bằng phương pháp đồng huấn luyện như sau:



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

Nội dung mục 3.2.3.1 và 3.2.3.2 dưới đây sẽ lần lượt trình bày các quá trình xây dựng mô hình học theo người dùng, xây dựng mô hình học theo sản phẩm từ tập dữ liệu huấn luyện. Trên cơ sở đó đề xuất kết hợp hai mô hình này trong một phương pháp đồng huấn luyện cho lọc cộng tác trong mục 3.2.3.3 và 3.2.3.4. Kết quả của quá trình đồng huấn luyện là ma trận đánh giá R đã tích hợp đầy đủ các giá trị đánh giá dự đoán sẽ được sử dụng để sinh tư vấn những sản phẩm phù hợp với người dùng hiện thời, nội dung này được đề cập trong mục 3.2.3.5.

3.2.3.1. Mô hình học theo người dùng

Như đã trình bày trong Mục 1.5.1, lọc cộng tác dựa vào bộ nhớ được thực hiện theo hai phương pháp chính: Lọc dựa vào người dùng (UserBased k-NN) và Lọc dựa vào sản phẩm (Item-Based k-NN) [1][10]. Tiếp cận phương pháp UserBased k-NN tính toán mức độ tương tự giữa người dùng $u_i \in U$ với tất cả những người dùng khác còn lại trên tập dữ liệu huấn luyện sẽ dẫn đến hai nhược điểm chính dưới đây.

Thứ nhất, nếu hai người dùng u_i, u_j có $|P_i \cap P_j|$ nhỏ nhưng có $r_{ix} = r_{jx}$ với mọi $p_x \in P_i \cap P_j$ thì hai người dùng này được xem là hoàn toàn giống nhau theo sở thích. Ví dụ Bảng 3.1 thể hiện một ma trận đánh giá trong hệ tư vấn cộng tác gồm 5

người dùng $U = \{u_1, u_2, u_3, u_4, u_5\}$ và 7 sản phẩm $P = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$. Từ Bảng 3.1, ta nhận thấy người dùng u_2 được xem là hoàn toàn tương tự với u_4 vì cả u_2, u_4 đều có đánh giá chung cho p_3, p_4 giống nhau ($r_{23} = r_{43} = 5$; $r_{24} = r_{44} = 5$). Kết quả là u_4 luôn là láng giềng của u_2 trong khi thực hiện dự đoán các sản phẩm mới cho u_2.

Bảng 3.1. Ma trận đánh giá của lọc cộng tác gồm 5 người dùng, 7 sản phẩm

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	0	3	0	3
u_2	5	0	5	5	4	0	0
u_3	4	0	0	4	3	4	3
u_4	0	3	5	5	0	5	0
u_5	?	5	?	?	0	4	4

Thứ hai, nếu hai người dùng u_i, u_j có $|P_i \cap P_j| = 0$ khi đó hai người dùng này được xem là hoàn toàn khác nhau theo sở thích (ví dụ u_2 và u_5 trong Bảng 3.1). Kết quả là các sản phẩm $p_x \in P_i \cap P_j = \emptyset$ sẽ không tham gia vào quá trình huấn luyện và dự đoán.

Trong đề xuất này, 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 [14], 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.*

Nhằm đơn giản hóa ký pháp trong trình bày, tập sinh S_i được xác định theo công thức (3.1) dưới đây.

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

Ví dụ chọn $\gamma = 3$, khi đó với người dùng u_1 của hệ đã cho trong Bảng 3.1 ta sẽ tìm được $S_1 = \{u_2, u_3\}$ vì cả u_2 và u_3 đều có 3 đánh giá chung với u_1. Tương tự như trên ta xác định được $S_2 = \{u_1, u_3\}, S_3 = \{u_1, u_2\}, S_4 = \emptyset, S_5 = \emptyset$.

Đối với các phương pháp tiếp cận theo người dùng trước đây, việc tính toán mức độ tương tự giữa người dùng $u_i \in U$ và người dùng $u_j \in U$ được thực hiện trên toàn bộ ma trận đánh giá để tìm ra tập K láng giềng của người dùng u_i. Phương pháp dự đoán các sản phẩm mới $p_x \in P$ được thực hiện dựa vào K láng giềng của người dùng u_i. Như vậy, tập K láng giềng của người dùng u_i là tập cố định không thay đổi trong quá trình huấn luyện [1][10] được xác định trên toàn bộ đánh giá.

Trong đề xuất này, mức độ tương tự của mỗi người dùng $u_i \in U$ và người dùng $u_j \in U$ chỉ được tính toán trên tập sinh $S_i \in U$ theo công thức (3.2). Điều này cho phép ta có thể ngăn ngừa được những cặp người dùng có $|P_i \cap P_j|$ nhỏ nhưng lại được đánh giá có tính tương tự cao theo phương pháp UserBased trước đây.

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

Như vậy mức độ tương tự giữa người dùng $u_i \in U$ với người dùng $u_j \in S_i$ xác định theo công thức (3.2) sẽ được dùng để tìm tập láng giềng cho người dùng u_i. Do tập sinh S_i của người dùng $u_i \in U$ luôn thay đổi theo mỗi vòng lặp của quá trình đồng huấn luyện, theo đó tập láng giềng của người dùng u_i là tập luôn thay đổi phụ thuộc vào tập sinh S_i của người dùng đó. Cụ thể, 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, $\beta \in [0,1]$.*

Nhằm đơn giản hóa ký pháp trong trình bày, tập láng giềng K_i được xác định theo công thức (3.3) dưới đây.

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

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).

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

Ví dụ tập với người dùng đã cho trong Bảng 3.1, ta tìm được $K_1 = \{u_3\}$, $K_2 = \{u_1\}$, $K_3 = \{u_1\}$. Khi đó, các giá trị dự đoán chắc chắn sẽ được điền cho u_1 là $r_{14} = 4, r_{16} = 4$. Giá trị dự đoán chắc chắn sẽ được điền cho u_2 là $r_{22} = 2,\ r_{27} = 3$. Giá trị dự đoán chắc chắn sẽ được điền cho u_3 là $r_{32} = 2,\ r_{33} = 5$. Kết quả đưa ra trong Bảng 3.2.

Rõ ràng, tập nhãn phân loại r_{ix} xác định theo công thức (3.4) nhỏ hơn rất nhiều so với tập r_{ix} xác định theo công thức (1.3). Tuy vậy, điều này sẽ được cải thiện dần thông qua việc quan sát các nhãn phân loại sinh ra trong quá trình học theo sản phẩm.

Bảng 3.2. Ma trận đánh giá ước lượng theo người dùng

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	**4**	3	**4**	3
u_2	5	**2**	5	5	4	0	**3**
u_3	4	**2**	**5**	4	3	4	3
u_4	0	3	5	5	0	5	0
u_5	?	5	?	?	0	4	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 $p_x \in P$ không dùng để xác định tập láng giềng K_x tác động trực

tiếp lên tư vấn như trong [1][10], mà chỉ để dùng vào việc xác định các nhãn phân loại chắc chắn r_{ix} cho sản phẩm p_x. Để thực hiện điều này, tác giả đưa ra 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.*

Nhằm đơn giản hóa ký pháp trong trình bày, tập sinh C_x được xác định theo công thức (3.5) dưới đây.

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

Ví dụ chọn $\gamma = 3$, khi đó với tập người dùng của hệ đã cho trong Bảng 3.1 ta sẽ tìm được $C_1 = \{p_5\}$, $C_2 = \emptyset, C_3 = \emptyset$, $C_4 = \emptyset$, $C_5 = \{p_1\}$, $C_6 = \emptyset$, $C_7 = \emptyset$. Tuy vậy, nếu việc huấn luyện theo sản phẩm được thực hiện sau quá trình huấn luyện theo người dùng như trong Bảng 3.2, ta sẽ xác định được:

$C_1 = \{p_2, p_3, p_4, p_5, p_7\}$, $C_2 = \{p_1, p_3, p_4, p_5, p_7\}$, $C_3 = \{p_1, p_2, p_4, p_5, p_7\}, C_4 = \{p_1, p_2, p_3, p_5, p_6, p_7\}$, $C_5 = \{p_1, p_2, p_3, p_4, p_7\}$, $C_6 = \{p_2, p_3, p_4, p_7\}$, $C_7 = \{p_1, p_2, p_3, p_4, p_5, p_6\}$.

Đối với các phương pháp tiếp cận theo sản phẩm trước đây, việc tính toán mức độ tương tự của các cặp sản phẩm $p_x \in P$ được thực hiện dựa trên toàn bộ ma trận đánh giá để tìm ra tập K láng giềng của sản phẩm p_x. Phương pháp dự đoán quan điểm của người dùng $u_i \in U$ với sản phẩm $p_x \in P$ được thực hiện dựa vào K láng giềng của sản phẩm p_x. Như vậy, K láng giềng của sản phẩm p_x là tập cố định không thay đổi trong quá trình huấn luyện

Trong đề xuất này, 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 $C_x \in P$ theo công thức (3.5). Điều này cho phép ngăn ngừa được các cặp sản phẩm có $|U_x \cap U_y|$ nhỏ nhưng lại được đánh giá có tính tương tự cao theo phương pháp ItemBased trước đây.

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

Như vậy mức độ tương tự giữa sản phẩm $p_x \in P$ với sản phẩm $p_y \in C_x$ theo công thức (3.6) sẽ được dùng để tìm tập láng giềng cho sản phẩm p_x. Do tập sinh C_x của sản phẩm $p_x \in P$ luôn thay đổi theo mỗi vòng lặp của quá trình đồng huấn luyện, theo đó tập láng giềng của sản phẩm $p_x \in P$ luôn thay đổi phụ thuộc vào tập sinh C_x của sản phẩm đó. Cụ thể, 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, $\beta \in [0,1]$.*

Nhằm đơn giản hóa ký pháp trong trình bày, tập láng giềng K_x được xác định theo công thức (3.7) dưới đây.

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

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).

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

Ví dụ với tập dữ liệu xuất phát như trong Bảng 3.2, ta tìm được $K_1 = \{p_4\}, K_2 = \{p_7\},\ K_3 = \{p_4\},\ K_4 = \{p_1\},\ K_5 = \{p_7\},\ K_6 = \{p_4\},\ K_7 = \{p_5\}$. Khi đó, các giá trị dự đoán chắc chắn sẽ được điền cho p_1 là $r_{41} = r_{44} = 5$, cho p_5 là $r_{55} = r_{57} = 4$, cho p_6 là $r_{26} = r_{24} = 5$. Khi đó, bảng giá trị đánh giá được thể hiện trong Bảng 3.3.

Bảng 3.3. Ma trận đánh ước lượng theo sản phẩm

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	**4**	3	**4**	3
u_2	5	**2**	5	5	4	**5**	**3**
u_3	4	**2**	**5**	4	3	4	3
u_4	**5**	3	5	5	0	5	0
u_5	?	5	?	?	**4**	4	4

Hai cơ chế quan sát theo sản phẩm và quan sát theo người dùng sẽ bổ sung nhãn phân loại chắc chắn cho các mẫu dữ liệu chưa có đánh giá theo mỗi bước thực hiện của quá trình đồng huấn luyện nhằm điền được tối đa các giá trị đánh giá chưa biết $r_{ix} = 0$ vào ma trận đánh giá. Muốn dự đoán thêm các nhãn phân loại tốt theo người dùng ta quan sát theo sản phẩm. Muốn dự đoán thêm các nhãn phân loại tốt theo sản phẩm ta quan sát theo người dùng. Quá trình học này được lặp lại luân phiên giữa 2 cơ chế quan sát theo người dùng và theo sản phẩm, theo một trong hai mô hình đồng huấn luyện đề xuất cho lọc cộng tác, đó là: Lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng CoTrainning-UserItem (Mục 3.2.3.3) và lọc cộng tác bằng phương pháp đồng huấn luyện theo sản phẩm CoTrainning-ItemUser (Mục 3.2.3.4) dưới đây. Điểm khác biệt cơ bản giữa hai phương pháp này là quá trình nào thực hiện trước, quá trình nào thực hiện sau trong cơ chế chuyển giao tri thức giữa các mô hình.

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

Phương pháp CoTrainning-UserItem được mô tả chi tiết trong Thuật toán 3.2 thực hiện thông qua t vòng lặp. Tại bước khởi tạo $t = 0$, ma trận dự đoán $R^{(0)} = \{r_{ix}^{(0)}\}$ được lấy bằng chính ma trận đánh giá ban đầu $R = \{r_{ix}\}$. Tại mỗi bước lặp, quá trình huấn luyện theo người dùng được thực hiện tuần tự theo các bước (2.1.a), (2.1.b),(2.1.c).

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

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

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.

Tại bước (2.1.a) ta cần xác định tập sinh $S_i^{(t)}$ cho người dùng u_i tại bước lặp thứ t theo công thức (3.1) và mức độ tương tự $u_{ij}^{(t)}$ giữa người $u_i \in U$ và người dùng $u_j \in S_i^{(t)}$ theo công thức (3.2). Tại bước (2.1.b), sử dụng $S_i^{(t)}, u_{ij}^{(t)}$ đã được xác định tại bước (2.1.a) ta xác định được $K_i^{(t)}$ là tập láng giềng của người dùng u_i tại bước lặp thứ t theo công thức (3.3). Tại bước (2.1.c), sử dụng $K_i^{(t)}$ đã xác định tại bước (2.1.b) ta dự đoán được $r_{ix}^{(t)}$ là quan điểm chắc chắn của người dùng u_i cho các sản phẩm p_x tại bước lặp thứ t theo công thức (3.4). Các giá trị $r_{ix}^{(t)}$ dự đoán theo người dùng tại bước lặp thứ t được bổ sung thêm vào quá trình huấn luyện theo sản phẩm tại bước 2.2.

Tại bước (2.2.a) ta cần xác định tập sinh $C_x^{(t)}$ của sản phẩm $p_x \in P$ theo công thức (3.5) và mức độ tương tự $p_{xy}^{(t)}$ giữa sản phẩm $p_x \in P$ và sản phẩm $p_y \in C_x^{(t)}$ theo công thức (3.6). Tại bước (2.2.b), sử dụng $C_x^{(t)}, p_{xy}^{(t)}$ đã được xác định tại bước (2.2.a) ta tìm được $K_x^{(t)}$ là tập láng giềng của sản phẩm p_x tại bước lặp thứ t theo công thức (3.7). Tại bước (2.2.c), sử dụng $K_x^{(t)}$ đã xác định tại bước (2.2.b) ta dự đoán được $r_{ix}^{(t)}$ là quan điểm chắc chắn của người dùng u_i cho các sản phẩm p_x tại bước lặp thứ t theo công thức (3.8).

Tại bước 2.3, bước lặp t được tăng lên 1 đơn vị và thực hiện quá trình đồng huấn luyện tiếp theo.

Mệnh đề 3.1 dưới đây sẽ làm sáng tỏ tính hội tụ của thuật toán.

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* $i = 1,2,\dots,N; x = 1,2,\dots,M.$

Chứng minh. Thực vậy, giả sử tại bước lặp thứ t $(t \geq 1)$, thuật toán *CoTraining-UserItem* có $r_{ix}^{(t)} = r_{ix}^{(t-1)}$, khi đó theo công thức(3.2) ta có:

$$u_{ij}^{(t)} = \frac{\sum_{p_x \in P_i^{(t)} \cap P_j^{(t)}} (r_{ix}^{(t)} - \overline{r_\imath^{(t)}})(r_{jx}^{(t)} - \overline{r_\jmath^{(t)}})}{\sqrt{\sum_{p_x \in P_i^{(t)} \cap P_j^{(t)}} (r_{ix}^{(t)} - \overline{r_\imath^{(t)}})^2 \sum_{P_i^{(t)} \cap P_j^{(t)}} (r_{jx}^{(t)} - \overline{r_\jmath^{(t)}})^2}}$$

$$= \frac{\sum_{p_x \in P_i^{(t-1)} \cap P_j^{(t-1)}} (r_{ix}^{(t-1)} - \overline{r_\imath^{(t-1)}})(r_{jx}^{(t-1)} - \overline{r_\jmath^{(t-1)}})}{\sqrt{\sum_{p_x \in P_i^{(t-1)} \cap P_j^{(t-1)}} (r_{ix}^{(t-1)} - \overline{r_\imath^{(t-1)}})^2 \sum_{P_i^{(t-1)} \cap P_j^{(t-1)}} (r_{jx}^{(t-1)} - \overline{r_\jmath^{(t-1)}})^2}}$$

$$u_{ij}^{(t)} = u_{ij}^{(t-1)} \leftrightarrow \begin{cases} S_i^{(t)} = S_i^{(t-1)} \\ K_i^{(t)} = K_i^{(t-1)} \end{cases}$$

Điều này chứng tỏ tại bước lặp thứ t, quá trình huấn luyện theo người dùng không bổ sung thêm được bất kỳ người dùng nào vào $S_i^{(t)}$ và $K_i^{(t)}$. Đây chính là

nguyên nhân quá trình huấn luyện theo người dùng không bổ sung được bất kỳ một nhãn phân loại $r_{ix}^{(t)}$ nào cho quá trình huấn luyện theo sản phẩm.

Đối với quá trình huấn luyện theo sản phẩm, theo công thức (3.6) ta cũng có:

$$p_{xy}^{(t)} = \frac{\sum_{u_i \in U_x^{(t)} \cap U_y^{(t)}} (r_{ix}^{(t)} - \overline{r_x^{(t)}})(r_{iy}^{(t)} - \overline{r_y^{(t)}})}{\sqrt{\sum_{u_i \in U_x^{(t)} \cap U_y^{(t)}} (r_{ix}^{(t)} - \overline{r_x^{(t)}})^2 \sum_{u_i \in U_x^{(t)} \cap U_y^{(t)}} (r_{iy}^{(t)} - \overline{r_y^{(t)}})^2}}$$

$$= \frac{\sum_{u_i \in U_x^{(t-1)} \cap U_y^{(t-1)}} (r_{ix}^{(t-1)} - \overline{r_x^{(t-1)}})(r_{iy}^{(t-1)} - \overline{r_y^{(t-1)}})}{\sqrt{\sum_{u_i \in U_x^{(t-1)} \cap U_y^{(t-1)}} (r_{ix}^{(t-1)} - \overline{r_x^{(t-1)}})^2 \sum_{u_i \in U_x^{(t-1)} \cap U_y^{(t-1)}} (r_{iy}^{(t-1)} - \overline{r_y^{(t-1)}})^2}}$$

$$p_{xy}^{(t)} = p_{xy}^{(t-1)} \leftrightarrow \begin{cases} C_x^{(t)} = C_x^{(t-1)} \\ K_x^{(t)} = K_{x-1}^{(t-1)} \end{cases}$$

Điều này chứng tỏ tại bước lặp thứ t, quá trình huấn luyện theo sản phẩm cũng không bổ sung thêm được bất kỳ sản phẩm nào vào $C_x^{(t)}$ và $K_x^{(t)}$. Đây chính là điểm hội tụ của thuật toán.

Trường hợp tốt nhất có thể đạt được là tất cả các giá trị $r_{ix} = 0$ đều được điền giá trị theo phương pháp *CoTraining-UserItem*. Định lý 3.1 dưới đây sẽ đưa ra điều kiện cần và đủ để các giá trị đánh giá đều được điền giá trị dự đoán.

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).*

Chứng minh (*điều kiện cần*). Giả sử với mọi người dùng $u_i \in U$ đều được dự đoán bởi giá trị $r_{ix} \neq 0$. Khi đó, ta cần chứng tỏ $\bigcup_{u_j \in K_i} P_j = P$.

Theo công thức (3.4) ta có:

$$r_{ix}^{(t)} = \overline{r_i^{(t)}} + \frac{\sum_{u_j \in K_i^{(t)}} \left(r_{jx}^{(t)} - \overline{r_j^{(t)}}\right) u_{ij}^{(t)}}{\sum_{u_j \in K_i^{(t)}} \left|u_{ij}^{(t)}\right|} \neq 0$$

$$\leftrightarrow \forall p_x \in P \ \exists u_j \in K_i^{(t)} | r_{jx}^{(t)} \neq 0$$

$$\leftrightarrow \bigcup_{u_j \in K_i} P_j = P$$

Ngược lại (*điều kiện đủ*), giả sử $\bigcup_{u_j \in K_i} P_j = P$. Khi đó ta cần chứng tỏ $r_{ix} \neq 0$ với $u_i \in U$.

Thực vậy, vì $\bigcup_{u_j \in K_i} P_j = P$ nên ta có :

$$\forall u_i \in U \quad \exists K_i^{(t)} : \bigcup_{u_j \in K_i^{(t)}} P_j = P$$

$$\leftrightarrow \forall p_y \in P \ \exists u_j \in K_i^{(t)} | r_{jx}^{(t)} \neq \emptyset$$

$$\leftrightarrow r_{ix}^{(t)} = \overline{r_i^{(t)}} + \frac{\sum_{u_j \in K_i^{(t)}} \left(r_{jx}^{(t)} - \overline{r_j^{(t)}}\right) u_{ij}^{(t)}}{\sum_{u_j \in K_i^{(t)}} \left|u_{ij}^{(t)}\right|} \neq 0$$

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

Phương pháp CoTraining-ItemUser mô tả chi tiết trong Thuật toán 3.3. Tại bước khởi tạo $t = 0$, ma trận dự đoán $R^{(0)} = (r_{ix}^{(0)})$ được lấy bằng chính ma trận đánh giá ban đầu $R = (r_{ix})$. Tại mỗi bước lặp, quá trình huấn luyện theo sản phẩm được thực hiện trước. Trong đó, bước (2.1.a) ta xác định định được tập sinh $C_x^{(t)}$ của sản phẩm $p_x \in P$ theo công thức (3.5), xác định mức độ tương tự $p_{xy}^{(t)}$ giữa sản phẩm $p_x \in P$ và sản phẩm $p_y \in C_x^{(t)}$ theo công thức (3.6). Sử dụng $C_x^{(t)}$, $p_{xy}^{(t)}$ đã được xác định tại bước (2.1.a) ta tìm được $K_x^{(t)}$ theo công thức (3.7) trong bước 2.1.b. Sử dụng $K_x^{(t)}$ đã xác định tại bước (2.1.b) ta dự đoán được $r_{ix}^{(t)}$ là quan điểm chắc chắn của người dùng u_i cho các sản phẩm p_x tại bước lặp thứ t theo công thức (3.8). Các giá trị $r_{ix}^{(t)}$ dự đoán theo sản phẩm tại bước lặp thứ t được bổ sung thêm vào quá trình huấn luyện theo người dùng tại bước 2.2.

Tại bước (2.2.a) ta cần xác định các tập $S_i^{(t)}$ theo công thức (3.1), $u_{ij}^{(t)}$ theo công thức (3.2). Sử dụng $S_i^{(t)}, u_{ij}^{(t)}$ đã được xác định tại bước (2.2.a) ta tìm được $K_i^{(t)}$ theo công thức (3.3). Sử dụng $K_i^{(t)}$ ta dự đoán được $r_{ix}^{(t)}$ là quan điểm chắc chắn của người dùng u_i cho các sản phẩm p_x tại bước lặp thứ t theo công thức (3.4). Tại bước 2.3, bước lặp t được tăng lên 1 đơn vị và thực hiện quá trình đồng huấn luyện tiếp theo.

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

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

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. Việc chứng minh mệnh đề 3.2, định lý 3.2 cũng được thực hiện tương tự như mệnh đề 3.1 và định lý 3.1.

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).*

3.2.3.5. Sinh tư vấn

Sau khi kết thúc quá trình đồng huấn luyện, hệ thống thu được ma trận dự đoán $R^{(t)}$ (Ma trận đánh giá R đã tích hợp đầy đủ các đánh giá dự đoán). Từ ma trận $R^{(t)}$ ta 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

Như đã đề cập trong Mục 3.1 đặt vấn đề, trong các hướng tiếp cận lọc kết hợp, 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.

Việc kết hợp đặc tính của lọc nội dung vào lọc cộng tác được thực hiện theo 2 cơ chế quan sát dữ liệu: 1) Quan sát theo người dùng cho phép hợp nhất hồ sơ người dùng của lọc nội dung vào ma trận đánh giá để thống nhất các mô hình dự đoán dựa vào người dùng; 2) Quan sát theo sản phẩm cho phép hợp nhất hồ sơ sản phẩm của lọc nội dung vào ma trận đánh giá để thống nhất các mô hình dự đoán dựa vào sản phẩm.

Trên cơ sở hợp nhất biểu diễn các giá trị đặc trưng nội dung vào ma trận đánh giá đề cập ở trên, luận án đề xuất phương pháp lọc kết hợp mới bằng phương pháp đồng huấn luyện. Phương pháp lọc kết hợp đề xuất phát triển trên cơ sở phương pháp đồng huấn luyện cho lọc cộng tác đề xuất trong Mục 3.3. Về cơ bản phương pháp lọc kết hợp bằng đồng huấn luyện đề xuất có cơ chế hoạt động tương tự với

phương pháp lọc cộng tác bằng đồng huấn luyện, nhưng có bổ sung giai đoạn xử lý hợp nhất biểu diễn các đặc trưng nội dung vào ma trận đánh giá khi quan sát dữ liệu theo người dùng và theo sản phẩm.

Nội dung trình bày sẽ được cấu trúc như sau: Mục 3.3.1 trình bày về việc kết hợp đặc tính của lọc nội dung vào lọc cộng tác dựa vào bộ nhớ; Mục 3.3.2 và Mục 3.3.3 lần lượt trình bày quá trình xây dựng mô hình học kết hợp theo người dùng, mô hình học kết hợp theo sản phẩm từ tập dữ liệu huấn luyện. Trên cơ sở đó đề xuất kết hợp hai mô hình này trong phương pháp đồng huấn luyện cho lọc kết hợp, nội dung này được trình bày trong Mục 3.3.4.

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

Không hạn chế tính tổng quát của bài toán tư vấn phát biểu trong Mục 1.3, ta giả thiết giá trị đánh giá của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$ được xác định theo công thức (3.9). Mỗi sản phẩm $p_x \in P$ biểu diễn thông qua $|C|$ đặc trưng nội dung $C = \{c_1, c_2, \dots, c_{|C|}\}$ được xác định theo công thức (3.10). Mỗi người dùng $u_i \in U$ biểu diễn thông qua $|T|$ đặc trưng nội dung $T = \{t_1, t_2, \dots, t_{|T|}\}$ được xác định theo công thức (3.11).

$$r_{ix} = \begin{cases} v, & \text{nếu người dùng } u_i \text{ đánh giá sản phẩm } p_x \text{ là } v \\ 0, & \text{nếu người dùng } u_i \text{ chưa đánh giá sản phẩm } p_x \end{cases} \quad (3.9)$$

$$c_{xs} = \begin{cases} 1, & \text{nếu sản phẩm } p_x \text{ có đặc trưng } c_s \\ 0, & \text{nếu sản phẩm } p_x \text{ không có đặc trưng } c_s \end{cases} \quad (3.10)$$

$$t_{iq} = \begin{cases} 1, & \text{nếu người dùng } u_i \text{ có đặc trưng } t_q \\ 0, & \text{nếu người dùng } u_i \text{ không có đặc trưng } t_q \end{cases} \quad (3.11)$$

Ví dụ với hệ gồm *3* người dùng $U = \{u_1, u_2, u_3\}$, *4* sản phẩm $P = \{p_1, p_2, p_3, p_4\}$. Trong đó, ma trận đánh giá R được cho trong Bảng 3.4, ma trận đặc trưng nội dung sản phẩm C được cho trong Bảng 3.5, ma trận đặc trưng nội dung người dùng T được cho trong Bảng 3.6.

Hệ tư vấn cộng tác được xây dựng dựa trên ma trận đánh giá R. Hệ tư vấn nội dung được xây dựng dựa trên ma trận các đặc trưng nội dung C và T. Hệ tư vấn lai xây dựng dựa trên cả ba ma trận R, C, T.

Bảng 3.4. Ma trận đánh giá R

	p_1	p_2	p_3	p_4
u_1	5	0	4	0
u_2	0	4	0	3
u_3	0	5	4	0

Bảng 3.5. Ma trận đặc trưng sản phẩm C

	c_1	c_2	c_3
p_1	1	0	1
p_2	1	1	0
p_3	1	0	1
p_4	0	1	1

Bảng 3.6. Ma trận đặc trưng người dùng T

	t_1	t_2	t_3	t_4
u_1	1	0	0	1
u_2	1	0	1	0
u_3	0	1	0	1

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

Để xây dựng được hồ sơ sử dụng các đặc trưng sản phẩm của người dùng, cần thực hiện hai nhiệm vụ: 1) Xác định tập sản phẩm người dùng đã từng truy cập hay sử dụng trong quá khứ; 2) Ước lượng trọng số mỗi đặc trưng nội dung sản phẩm trong hồ sơ người dùng [48][114]. 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 $u_i \in U$.

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

Khi đó, P_i chính là tập sản phẩm người dùng đã từng truy cập trong quá khứ được các phương pháp tư vấn theo nội dung sử dụng trong khi xây dựng hồ sơ người dùng. Vấn đề còn lại là làm thế nào ta ước lượng được trọng số mỗi đặc trưng $c_s \in C$ đối với mỗi hồ sơ người dùng $u_i \in U$.

Gọi $Item(i,s)$ 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). Khi đó, $|Item(i,s)|$ chính là số lần người dùng $u_i \in U$ sử dụng các sản phẩm $p_x \in P$ chứa đựng đặc trưng $c_s \in C$ trong quá khứ.

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

Dựa trên P_i và $Item(i,s)$ 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. Phương pháp phổ dụng nhất được sử dụng trong xây dựng hồ sơ người dùng là kỹ thuật tf-idf [48][114], khi đó giá trị w_{is} là một số thực trải đều trong khoảng [0,1]. Tuy nhiên, trong khi quan sát bài toán tư vấn cộng tác tác giả nhận thấy bản thân nó đã tồn tại một phép đánh giá tự nhiên của người dùng đối với sản phẩm thông qua giá trị đánh giá r_{ix}. Giá trị r_{ix} phản ánh mức độ ưa thích của người dùng sau khi đã sử dụng sản phẩm và đưa ra quan điểm của mình đối với sản phẩm. Ví dụ với hệ tư vấn phim [14][13], giá trị $r_{ix} = 1,2,3,4,5$ được hiểu theo các mức quan điểm "*rất tồi*", "*tồi*", "*bình thường*", "*hay*", "*rất hay*". Chính vì lý do đó, tác giả mong muốn có được một phép trích chọn đặc trưng có cùng mức độ đánh giá tự nhiên của r_{ix}.

Để thực hiện ý tưởng nêu trên, tác giả thực hiện quan sát trên tập $Item(i,s)$. Nếu giá trị $|Item(i,s)|$ vượt quá một ngưỡng θ nào đó thì trọng số đặc trưng nội dung sản phẩm $c_s \in C$ đối với người dùng $u_i \in U$ là w_{is} được tính bằng trung bình cộng của tất cả các giá trị đánh giá. Trường hợp $|Item(i,s)|$ có giá trị bé hơn θ, giá trị w_{is} được tính bằng tổng của tất cả các giá trị đánh giá chia cho θ. Trong thực nghiệm, tác giả tính toán được số lượng trung bình của tất cả người dùng $u_i \in U$ đã đánh giá các sản phẩm $p_x \in P$. Sau đó chọn θ tương đương với $2/3$ số lượng trung bình các đánh giá của tập người dùng $u_i \in U$ đã đánh giá sản phẩm $p_x \in P$ chứa đựng đặc trưng $c_s \in C$. Bằng cách này ta có thể hạn chế được một số đặc trưng nội dung ít được người dùng quan tâm nhưng vẫn được đánh giá với trọng số cao.

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

Giá trị w_{is} được ước lượng theo (3.14) phản ánh quan điểm của người dùng $u_i \in U$ đối với các đặc trưng nội dung sản phẩm $c_s \in C$ trong quá khứ. 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. Dựa trên nhận xét này, tác giả hợp nhất ma trận đánh giá của lọc cộng tác và hồ sơ người dùng của lọc nội dung thành mô hình biểu diễn hợp nhất giữa đánh giá người dùng của lọc cộng tác với các đặc trưng sản phẩm của lọc nội dung. 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.

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

Ví dụ với hệ có ma trận đánh giá theo Bảng 3.4, ma trận đặc trưng sản phẩm theo Bảng 3.5, chọn $\theta = 2$, khi đó ta sẽ tính toán được tập hồ sơ người dùng $\{w_{is} : u_i \in U, c_s \in C\}$ trong Bảng 3.7 và ma trận đánh giá mở rộng về phía sản phẩm theo (3.15) trong Bảng 3.8.

Bảng 3.7. Ma trận hồ sơ người dùng ($\boldsymbol{w_{is}}$)

	c_1	c_2	c_3
u_1	4	0	4
u_2	2	3	1
u_3	4	2	2

Bảng 3.8. Ma trận đánh giá mở rộng ($\boldsymbol{r_{ix}}$) theo hồ sơ người dùng

	p_1	p_2	p_3	p_4	c_1	c_2	c_3
u_1	5	0	4	0	4	0	4
u_2	0	4	0	3	2	3	1
u_3	0	5	4	0	4	2	2

Hệ tư vấn được xác định theo (3.15) đã tích hợp đầy đủ đánh giá người dùng và trọng số các đặc trưng sản phẩm. Chính vì vậy, các phương pháp tư vấn kết hợp dựa vào người dùng đều có thể dễ dàng triển khai trên ma trận đánh giá mở rộng theo hồ sơ người dùng [47][55]. Do tính chất thưa thớt của ma trận đánh giá ban đầu làm cho ma trận đánh giá mở rộng theo hồ sơ người dùng cũng thưa thớt. Chính vì vậy, các phương pháp tư vấn chỉ dựa vào (3.15) đều cho lại kết quả không cao. Vấn đề này sẽ được giải quyết trong Mục 3.3.4.

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

Tương tự như hồ sơ người dùng, hồ sơ sản phẩm lưu trữ lại dấu vết các đặc trưng nội dung người dùng đã từng sử dụng sản phẩm. Để xây dựng được hồ sơ sản phẩm, cần thực hiện xác định tập người dùng đã từng sử dụng sản phẩm trong quá khứ và ước lượng trọng số mỗi đặc trưng nội dung người dùng trong hồ sơ sản phẩm [47]. 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 $p_x \in P$. Khi đó, U_x chính là tập người dùng cần được lưu lại các giá trị đặc trưng nội dung trong hồ sơ sản phẩm. Vấn đề còn lại là làm thế nào ta ước lượng được trọng số mỗi đặc trưng $t_q \in T$ đối với mỗi hồ sơ 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). Khi đó, $|User(x, q)|$ chính là số lần sản phẩm $p_x \in P$ được tập người dùng $u_i \in$ U có đặc trưng nội dung $t_q \in T$ sử dụng trong quá khứ.

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

Giống như người dùng, bản thân các sản phẩm cũng đã tồn tại một phép đánh giá tự nhiên của tập người dùng đối với sản phẩm thông qua giá trị đánh giá r_{ix}. Do vậy, 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}. Để thực hiện điều này, tác giả tiến hành quan sát trên tập $User(x, q)$. Nếu giá trị $|User(x, q)|$ vượt quá một ngưỡng θ

nào đó thì trọng số đặc trưng nội dung người dùng $t_q \in T$ đối với sản phẩm $p_x \in P$ là v_{qx} được tính bằng trung bình cộng của tất cả các giá trị đánh giá. Trường hợp $|User(x,q)|$ có giá trị bé hơn θ, giá trị v_{qx} được tính bằng tổng của tất cả các giá trị đánh giá chia cho θ.

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

Giá trị v_{qx} được ước lượng theo (3.18) biểu diễn hồ sơ sản phẩm $p_x \in P$ đã được tập những người dùng $u_i \in U$ chứa đựng đặc trưng $t_q \in T$ sử dụng. Vì vậy, ta có thể xem mỗi đặc trưng nội dung người dùng đóng vai trò như một người dùng phụ bổ sung vào tập người dùng. Dựa trên nhận xét này, tác giả hợp nhất ma trận đánh giá của lọc cộng tác và hồ sơ sản phẩm của lọc nội dung thành mô hình biểu diễn hợp nhất giữa đánh giá sản phẩm của lọc cộng tác với các đặc trưng người dùng của lọc nội dung. 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.

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

Ví dụ với hệ có ma trận đánh giá theo Bảng 3.4, ma trận đặc trưng người dùng theo Bảng 3.6, chọn $\theta = 2$, khi đó ta sẽ tính toán được tập hồ sơ sản phẩm $\{v_{qx} : p_x \in P, t_q \in T\}$ trong Bảng 3.9 và ma trận đánh giá mở rộng về phía người dùng theo (3.15) trong Bảng 3.8.

Bảng 3.9. Ma trận hồ sơ sản phẩm (v_{qx})

	p_1	p_2	p_3	p_4
t_1	2	2	2	1
t_2	0	0	2	0
t_3	0	2	0	1
t_4	2	2	4	0

Bảng 3.10. Ma trận đánh giá mở rộng ($\boldsymbol{r_{ix}}$) theo hồ sơ sản phẩm

	p_1	p_2	p_3	p_4
u_1	5	0	4	0
u_2	0	4	0	3
u_3	0	5	4	0
t_1	2	2	2	1
t_2	0	0	2	0
t_3	0	2	0	1
t_4	2	2	4	0

Hệ tư vấn được xác định theo (3.19) đã tích hợp đầy đủ đánh giá sản phẩm và trọng số các đặc trưng người dùng. Chính vì vậy, các phương pháp tư vấn kết hợp theo sản phẩm đều có thể dễ dàng triển khai trên ma trận đánh giá mở rộng theo hồ sơ sản phẩm [47][13]. Do tính chất thưa thớt của ma trận đánh giá ban đầu làm cho ma trận đánh giá mở rộng theo hồ sơ sản phẩm cũng thưa thớt. Chính vì vậy, các phương pháp tư vấn dựa vào (3.19) đều cho lại kết quả không cao. Vấn đề này sẽ được giải quyết trong Mục 3.3.4.

3.3.2. Mô hình học kết hợp 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).

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

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

S_i được xác định theo (3.20) là tập người dùng $u_j \in U$ có số lượng đánh giá giao nhau với người dùng u_i ít nhất là θ_1 sản phẩm và số lượng các đặc trưng sản phẩm giao nhau ít nhất là θ_2. Hai hằng số nguyên dương θ_1 và θ_2 được chọn đủ lớn trong tập dữ liệu huấn luyện để S_i không còn là tập dữ liệu thưa. Dựa vào S_i và độ

tương quan Pearson [114][14], 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).

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

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

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

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).

$$H_i = P_i \cup C_i \quad (3.25)$$

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

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

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

Rõ ràng, a_{ij} được xác định trên S_i theo (3.22) chính xác hơn so với a_{ij} được xác định trên toàn bộ tập người dùng U trong tập dữ liệu huấn luyện vì S_i chiếu lên các cột sản phẩm không phải là tập dữ liệu thưa. Giá trị b_{ij} được xác định trên S_i theo

(3.23) chính xác hơn so với b_{ij} được xác định trên toàn bộ đặc trưng sản phẩm C vì S_i chiếu lên các cột đặc trưng sản phẩm cũng không phải là tập dữ liệu thưa. Giá trị u_{ij} được xác định theo (3.24) tin cậy hơn so với u_{ij} xác định trên toàn bộ tập người dùng vì S_i không phải là tập dữ liệu thưa trên toàn bộ $P \cup C$. Hơn thế nữa, hai người dùng u_i, u_j có mức độ tương tự theo đánh giá người dùng và tương tự theo hồ sơ người dùng phải vượt quá một ngưỡng α nào đó. Ngưỡng α được xác định thông qua kiểm nghiệm. Trong thực nghiệm tác giả chọn $\alpha = 0.9$ để có được kết quả tốt nhất.

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) [15][10].

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

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

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).

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

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

S_x được xác định theo (3.31) là tập sản phẩm $p_y \in P$ có số lượng người dùng đánh giá giao nhau với sản phẩm p_x ít nhất là γ_1 và số lượng các đặc trưng người dùng giao nhau ít nhất là γ_2. Hai hằng số nguyên dương γ_1 và γ_2 được chọn đủ lớn trong tập dữ liệu huấn luyện để S_x không còn là tập dữ liệu thưa. 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).

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

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

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

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.

$$H_x = U_x \cup T_x \tag{3.36}$$

$$\bar{r_x} = \frac{1}{|U_x \cap U_y|} \sum_{i \in U_x \cap U_y} r_{ix} \tag{3.37}$$

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

$$\bar{\bar{r_x}} = \frac{1}{|H_x \cap H_y|} \sum_{i \in H_x \cap H_y} r_{ix} \tag{3.39}$$

Rõ ràng, a_{xy} được xác định trên S_x theo (3.33) chính xác hơn so với a_{xy} được xác định trên toàn bộ tập sản phẩm P trong tập dữ liệu huấn luyện vì S_x chọn trên các hàng người dùng không phải là tập dữ liệu thưa. Giá trị b_{xy} được xác định trên S_x theo (3.34) chính xác hơn so với b_{xy} được xác định trên toàn bộ tập đặc trưng người dùng T vì S_x chọn trên các hàng đặc trưng người dùng cũng không phải là tập dữ liệu thưa. Giá trị p_{xy} được xác định theo (3.35) tin cậy hơn so với p_{xy} xác định trên toàn bộ tập sản phẩm và đặc trưng người dùng vì S_x không phải là tập dữ liệu thưa trên toàn bộ $U \cup T$. Hơn thế nữa, hai sản phẩm p_x, p_y có mức độ tương tự theo đánh giá sản phẩm và tương tự theo hồ sơ sản phẩm phải vượt quá một ngưỡng α nào đó. Ngưỡng α được xác định thông qua kiểm nghiệm. Trong thực nghiệm tác giả chọn $\alpha = 0.9$ để có được kết quả tốt nhất.

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) [13][10].

$$K_x = \{y \in S_x : p_{xy} > \alpha\} \tag{3.40}$$

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

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

Trong mục này, luận án đưa ra đề xuất một phương pháp lọc kết hợp mới bằng mô hình đồng huấn luyện. Mô hình đồng huấn luyện đề xuất cho lọc kết hợp phát triển từ mô hình đồng huấn luyện cho lọc cộng tác trình bày trong Mục 3.2.3.

Theo đó, mô hình đồng huấn luyện cho lọc kết hợp đề xuất sẽ học các mẫu dữ liệu độc lập khi quan sát theo người dùng và theo sản phẩm để gán nhãn cho các mẫu dữ liệu chưa biết, trong trường hợp này là dự đoán đánh giá của người dùng với sản phẩm mới. Khi quan sát dữ liệu theo người dùng, mô hình đồng huấn luyện thực hiện học bán giám sát theo các đánh giá người dùng với sản phẩm và tập đặc trưng sản phẩm, điều này cho phép ta phát hiện những sản phẩm mới phù hợp nhất đối với mỗi người dùng. Khi quan sát dữ liệu theo sản phẩm, mô hình đồng huấn luyện thực hiện học bán giám sát theo các đánh giá sản phẩm bởi người dùng và tập đặc trưng người dùng, điều này cho phép ta phát hiện những người dùng mới phù hợp nhất đối với mỗi sản phẩm. Hai quá trình học này được lặp lại luân phiên giữa 2 cơ chế quan sát dữ liệu theo người dùng và theo sản phẩm đến khi thỏa mãn điều kiện các mẫu dữ liệu đều được gán nhãn hoặc số vòng lặp đạt đến ngưỡng xác định. Từ đó phương pháp đề xuất góp phần giải quyết ảnh hưởng của vấn đề dữ liệu thưa của ma trận đánh giá. Thuật toán được mô tả chi tiết như trong Thuật toán 3.4.

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

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

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

Tại bước (2.1), quá trình huấn luyện kết hợp theo người dùng sẽ thực hiện học bán giám sát theo các đánh giá người dùng với sản phẩm và tập đặc trưng sản phẩm, được thực hiện tuần tự theo các bước (2.1.a), (2.1.b), (2.1.c), (2.1.d), (2.1.e), (2.1.f). Tại bước (2.1.a) ta xác định được $w_{is}^{(t)}$ phản ánh quan điểm của tập người dùng $u_i \in U$ đối với sản phẩm $c_s \in C$ của vòng lặp thứ (t) theo công thức (3.14). Sử dụng $w_{is}^{(t)}$, tại bước (2.1.b) ta xây dựng được ma trận đánh giá mở rộng theo hồ sơ người dùng của vòng lặp thứ (t) theo công thức (3.15). Dựa vào kết quả của bước (2.1.b), tại bước (2.1.c) ta xác định được tập $S_i^{(t)}$ là tập dữ liệu không thưa đối với người dùng $u_i \in U$ của vòng lặp thứ (t) theo công thức (3.20). Sử dụng $S_i^{(t)}$, bước (2.1.d) ta xác định được $u_{ij}^{(t)}$ là mức độ tương tự giữa các cặp người dùng $u_i, u_j \in U$ trên cả tập đánh giá của người dùng với sản phẩm và tập đặc trưng người dùng của vòng lặp thứ (t) theo công thức (3.24). Sau khi tính toán được $u_{ij}^{(t)}$, tại bước (2.1.e) ta xác định được $K_i^{(t)}$ là tập láng giềng của người dùng u_i của vòng lặp thứ (t) theo công thức (3.29). Cuối cùng, tại bước (2.1.f) ta dự đoán được giá trị $r_{ix}^{(t)}$ phản ánh mức độ phù hợp của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$ tại vòng lặp thứ (t). Các giá trị $r_{ix}^{(t)}$ dự đoán được tại vòng lặp thứ (t) sẽ được cập nhật lại trong ma trận đánh giá mở rộng $R^{(t)}$ và chuyển giao cho quá trình huấn luyện kết hợp theo sản phẩm.

Tại bước (2.2), quá trình huấn luyện kết hợp theo sản phẩm sẽ thực hiện học bán giám sát theo các đánh giá sản phẩm bởi người dùng và tập đặc trưng người dùng, được thực hiện tuần tự theo các bước (2.2.a), (2.2.b), (2.2.c), (2.2.d), (2.2.e), (2.2.f). Tại bước (2.2.a) ta xác định được $v_{qx}^{(t)}$ phản ánh quan điểm của tập người dùng có đặc trưng nội dung $t_q \in T$ đối với sản phẩm $p_x \in P$ của vòng lặp thứ (t) theo công thức (3.18). Sử dụng $v_{qx}^{(t)}$, tại bước (2.2.b) ta xây dựng được ma trận đánh giá mở rộng theo hồ sơ sản phẩm của vòng lặp thứ (t) theo công thức (3.19). Dựa vào kết quả của bước (2.2.b), tại bước (2.2.c) ta xác định được tập $S_x^{(t)}$ là tập dữ liệu không

thưa đối với sản phẩm $p_x \in P$ của vòng lặp thứ (t) theo công thức (3.31). Sử dụng $S_x^{(t)}$, bước (2.2.d) ta xác định được $p_{xy}^{(t)}$ là mức độ tương tự giữa các cặp sản phẩm $p_x, p_y \in P$ trên cả tập đánh giá của người dùng và đặc trưng người dùng với sản phẩm tại vòng lặp thứ (t) theo công thức (3.35). Sau khi tính toán được $p_{xy}^{(t)}$, tại bước (2.2.e) ta xác định được $K_x^{(t)}$ là tập láng giềng của sản phẩm p_x tại vòng lặp thứ (t) theo công thức (3.40). Cuối cùng, tại bước (2.2.f) ta dự đoán được giá trị $r_{ix}^{(t)}$ phản ánh mức độ phù hợp của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$ tại vòng lặp thứ (t). Các giá trị $r_{ix}^{(t)}$ dự đoán được tại vòng lặp thứ (t) sẽ được cập nhật lại trong ma trận đánh giá mở rộng $R^{(t)}$ và chuyển giao cho quá trình huấn luyện kết hợp theo người dùng tại bước lặp tiếp theo của thuật toán.

Tại bước (2.3), số lượng vòng lặp (t) được tăng lên 1 đơn vị và thuật toán tiếp tục lặp lại quá trình đồng huấn luyện tiếp theo. Thuật toán sẽ hội tụ tại vòng lặp thứ (t) có $u_{ij}^{(t)} = u_{ij}^{(t-1)}$ và $p_{xy}^{(t)} = p_{xy}^{(t-1)}$. Tại bước 3 của thuật toán, quá trình tạo nên tư vấn được thực hiện đơn giản bằng cách sắp xếp theo thứ tự giảm dần các giá trị dự đoán $r_{ix}^{(t)}$, sau đó chọn K sản phẩm $p_x \in P$ có giá trị $r_{ix}^{(t)}$ lớn nhất tư vấn cho người dùng hiện thời.

3.4. Thực nghiệm và kết quả

Nội dung phần này trình bày đánh giá hiệu quả về độ chính xác của phương pháp đề xuất trong sự so sánh với các phương pháp tư vấn cơ sở. Có hai đề xuất được trình bày trong chương này: 1) Đề xuất phương pháp lọc cộng tác bằng đồng huấn luyện; 2) Đề xuất phương pháp lọc kết hợp bằng đồng huấn luyện, trong đó phương pháp lọc kết hợp phát triển từ phương pháp lọc cộng tác đề xuất. Do vậy để tường minh kết quả, nội dung thực nghiệm sẽ được chia thành 2 phần: Mục 3.4.1 trình bày nội dung thực nghiệm và đánh giá kết quả của phương pháp lọc cộng tác bằng đồng huấn luyện và Mục 3.4.2 trình bày nội dung thực nghiệm và đánh giá kết quả của phương pháp lọc kết hợp bằng đồng huấn luyện

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 các bộ dữ liệu MovieLens của nhóm nghiên cứu GroupLens thuộc trường đại học Minnesota. Bộ dữ liệu thứ nhất *MovieLens-100K* bao gồm 100.000 đánh giá của 943 người dùng cho 1682 phim. Giá trị đánh giá được thực hiện từ 1 đến 5. Mức độ thưa thớt dữ liệu đánh giá là 93.7%. Bộ dữ liệu thứ hai *MovieLens-1M* bao gồm 1000.000 đánh giá của 6000 người dùng cho 4000 phim. Bộ dữ liệu thứ ba *MovieLens-10M* bao gồm 10.000.000 đánh giá của 72000 người dùng với 10.000 bộ phim (http://www.grouplens.org/).

3.4.1.2. Cài đặt thực nghiệm

Độ đo đánh giá

Hai nhiệm vụ chính của hệ tư vấn là dự đoán đánh giá và tư vấn danh sách ngắn các sản phẩm cho người dùng hiện thời. Để đánh giá hiệu quả của đánh giá dự đoán, các độ đo thường được sử dụng là $MAE, RMSE, MPE$. Để đánh giá hiệu quả tư vấn danh sách sản phẩm, các độ đo điển hình được sử dụng là $Precision@N$, $Recall@N$ và $MAP@N$. Trong hai hướng đánh giá như vậy, trong chương này luận án tập trung vào nhóm độ đo đánh giá dự đoán cho bài toán tư vấn để đánh giá hiệu quả của phương pháp đề xuất trong sự so sánh với các phương pháp lọc cộng tác cơ sở. Trong đó, độ đo được sử dụng phổ biến để thực nghiệm là $MAE, RMSE$, chi tiết về các độ đo này được trình bày trong Mục 1.6.2 của luận án. Giá trị $MAE, RMSE$ càng nhỏ thể hiện thuật toán tư vấn có độ chính xác càng cao.

Phương pháp thực nghiệm

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn, tác giả sử dụng lý thuyết về phương pháp đánh giá hệ thống được trình bày trong Mục 1.6.1 của luận án. Trong đó, 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.

Để thực nghiệm khả năng của phương pháp mới đề xuất so với những phương pháp khác trong trường hợp có ít dữ liệu, tác giả thay đổi số lượng đánh giá của mỗi người dùng trong tập kiểm tra sao cho số lượng đánh giá đã biết lần lượt là 5, 10 và 20, phần còn lại là những đánh giá cần dự đoán. Chọn $\beta = 0.8$ và $\gamma = 4, 7, 14$ ứng với các tập dữ liệu kiểm tra biết trước 5, 10, 20 đánh giá. Điều này có nghĩa, việc tính toán mức độ tương tự giữa các cặp người dùng hoặc các cặp sản phẩm chỉ thực hiện trên các cặp người dùng hoặc sản phẩm giao nhau trên 2/3 sản phẩm hoặc người dùng cùng đánh giá. Tập láng giềng giữa các cặp người dùng hoặc sản phẩm chỉ được lấy trên tập sinh có mức độ tương tự là 0.8 (rất giống nhau). Số bước lặp giới hạn cho các phương pháp đồng huấn luyện tương ứng là 48, 53, 55 với lần lượt 3 tập dữ liệu thực nghiệm, nhằm đảm bảo các mẫu dữ liệu chưa có nhãn gần như được gán nhãn đầy đủ.

Các phương pháp lọc cộng tác được sử dụng trong thực nghiệm

- *Phương pháp UserBased*: sử dụng độ tương quan Pearson [10][15]. Đây là phương pháp lọc cộng tác theo bộ nhớ dựa vào người dùng đã được trình bày trong Mục 1.5.1.1.
- *Phương pháp ItemBased*: sử dụng độ tương quan Pearson [10][13]. Đây là phương pháp lọc cộng tác dựa trên sản phẩm đã được trình bày trong Mục 1.5.1.1.
- *Phương pháp CoTraining-UserItem*: đề xuất trong Mục 3.3.3.1 của luận án.
- *Phương pháp CoTraining-ItemUser*: đề xuất trong Mục 3.3.3.2 của luận án.

Để thực nghiệm các phương pháp này, tác giả sử dụng máy tính có cấu hình: Intel Core i7-3770 CPU và 8GB RAM. Các thuật toán tư vấn cơ sở đã có được thực nghiệm dựa vào bộ thư viện CARSKIT [103], các thuật toán tư vấn đề xuất do tác giả tự cài đặt và thực nghiệm theo cùng chiến lược thực nghiệm với phương pháp cơ sở. Việc thực nghiệm được thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.

3.4.1.3. Kết quả kiểm nghiệm

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.

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 3.13. 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**

Trong trường hợp dữ liệu tương đối đầy đủ, cụ thể là khi biết trước nhiều đánh giá của người dùng trong tập kiểm tra, phương pháp *CoTraining-UserItem* và *CoTraining-ItemUser* cho lại kết quả tương đương nhau. Tuy nhiên, khi dữ liệu ít đi, cụ thể là khi chỉ biết trước 5 hoặc 10 đánh giá của người dùng kiểm tra thì trong đa số trường hợp, *CoTraining-ItemUser* cho sai số MAE, RMSE nhỏ hơn so với *CoTraining-UserItem*. Lý do chủ yếu là do lực lượng của tập C_x xác định theo (3.5) lớn hơn lực lượng của tập S_i xác định theo (3.1). Điều này cho phép dự đoán các nhãn phân loại bổ sung vào quá trình huấn luyện theo người dùng tốt hơn.

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

Thuật toán đồng huấn luyện cho lọc kết hợp được thực nghiệm trên bộ dữ liệu MovieLens của nhóm nghiên cứu GroupLens thuộc trường đại học Minnesota. Tập dữ liệu MovieLens có ba lựa chọn với kích thước khác nhau lần lượt là: *MovieLens 100KB, MovieLens 1MB* và *MovieLens 10MB*. Trong đó, tập dữ liệu *MovieLens 100KB* là tập con của tập *MovieLens 1MB*. Tập dữ liệu *MovieLens 1MB* cung cấp đầy đủ tập đặc trưng sản phẩm và người dùng kèm theo tập đánh giá người dùng. Tập dữ liệu *MovieLens 10M* tuy lớn nhưng không cung cấp tập đặc trưng người dùng và tập đặc trưng sản phẩm. Chính vì vậy, 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. Giá trị đánh giá được thực hiện từ 1 đến 5. Mức độ thưa thớt dữ liệu đánh giá là 95.83%. Dữ liệu cụ thể được cung cấp trong các file sau:

- *u.data*: lưu trữ đầy đủ 1MB đánh giá của 6040 người dùng cho 3952 phim. Mỗi người dùng đánh giá ít nhất 20 phim. Mỗi hàng đều có cùng cấu trúc: user id | item id | rating | timestamp.
- *u.info*: File lưu số lượng người dùng, số lượng sản phẩm, số lượng xếp hạng của tập dữ liệu. File u.item lưu thông tin về phim.

- *u.genre*: File lưu danh sách 18 thể loại phim khác nhau. Đây là tập đặc trưng nội dung sản phẩm được dùng trong thực nghiệm phương pháp đề xuất. Ngoài ra, ứng với mỗi phim tách trong IMDB để lấy tập đặc trưng nước sản xuất, hãng phim, đạo diễn, diễn viên chính để làm tập đặc trưng phim.
- *u.user*: File lưu thông tin về những người dùng. Các hàng có cấu trúc chung : user id | age | gender | occupation | zip code. User id được sử dụng trong tập dữ liệu u.data.
- *u.occupation*: File lưu danh sách các nghề nghiệp. Đây là tập đặc trưng nội dung người dùng được dùng trong thực nghiệm phương pháp đề xuất.

3.4.2.2. Cài đặt thực nghiệm

Độ đo đánh giá

Để đánh giá hiệu quả về độ chính xác của phương pháp đề xuất trong sự so sánh với các phương pháp lọc cộng tác cơ sở, độ đo sử dụng để kiểm nghiệm là $MAE, RMSE$ được trình bày trong Mục 1.6.2 của luận án. Lý do cho việc lựa chọn này được lý luận tương tự Mục 3.5.1.2. Giá trị $MAE, RMSE$ càng nhỏ thể hiện thuật toán tư vấn có độ chính xác càng cao.

Phương pháp thực nghiệm

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn, tác giả sử dụng lý thuyết về phương pháp đánh giá hệ thống được trình bày trong Mục 1.6.1 của luận án. Trong đó, 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. Tập test1.inp, test2.inp, test3.inp có số đánh giá giá biết trước lần lượt của mỗi người dùng là 5, 10, 15 tương ứng với trường hợp dữ liệu huấn luyện thưa. Tập test4.inp có số đánh giá giá biết trước là 20 tương ứng với trường hợp dữ liệu huấn luyện

tương đối đầy đủ. Chọn $\theta = 4, 8, 12, 15$ ứng với mỗi bộ test theo thứ tự để xác định xác định w_{is}, v_{qx} theo công thức (3.14), (3.18). Chọn $\theta_1 = 4, 8, 12, 15$ (cho mỗi tập dữ liệu theo thứ tự), $\theta_2 = 10$ và $\alpha = 0.9$ (cho tất cả các tập dữ liệu kiểm tra) để xác định S_i, u_{ij}, K_i theo công thức (3.20), (3.24), (3.29), và S_x, p_{xy}, K_x theo công thức (3.31), (3.35), (3.40).

Các phương pháp tư vấn được sử dụng trong thực nghiệm

- Phương pháp tư vấn cộng tác dựa vào người dùng sử dụng độ tương quan Pearson (ký hiệu là *CF-UserBased*) [15][10].
- Phương pháp tư vấn cộng tác dựa vào sản phẩm sử dụng độ tương quan Pearson (ký hiệu là *CF-ItemBased*) [13][10].
- Phương pháp tư vấn nội dung dựa vào hồ sơ người dùng sử dụng độ tương quan Pearson (ký hiệu là *CBF-UserBased*) [48].
- Phương pháp tư vấn nội dung dựa vào hồ sơ sản phẩm sử dụng độ tương quan Pearson (ký hiệu là *CBF-ItemBased*) [41].
- Phương pháp tư vấn kết hợp dựa vào người dùng và tập đặc trưng sản phẩm sử dụng độ tương quan Pearson (ký hiệu là *Hybrid-UserBased*). Đây là phương pháp tư vấn kết hợp dựa vào độ tương quan Pearson được đề xuất theo công thức (3.24).
- Phương pháp tư vấn kết hợp dựa theo sản phẩm và tập đặc trưng người dùng sử dụng độ tương quan Pearson (ký hiệu là *Hybrid-ItemBased*). Đây là phương pháp tư vấn kết hợp dựa vào độ tương quan Pearson được đề xuất theo công thức (3.35).
- Phương pháp lọc kết hợp bằng đồng huấn luyện đề xuất trong Mục 3.4.4 (Ký hiệu là CoTraining- HybridFiltering).

Để thực nghiệm các phương pháp này, tác giả sử dụng máy tính có cấu hình: Intel Core i7-3770 CPU và 8GB RAM. Các thuật toán tư vấn theo ngữ cảnh cơ sở đã có được thực nghiệm dựa vào bộ thư viện CARSKIT [103], các thuật toán tư vấn

đề xuất do tác giả tự cài đặt và thực nghiệm theo cùng chiến lược thực nghiệm với phương pháp cơ sở và dùng cùng các độ đo đánh giá dự đoán phổ biến (MAE, RMSE) để so sánh. Để đảm bảo đánh giá công bằng phương pháp đề xuất so với các phương pháp cơ sở, tác giả thực hiện 10 lần và lấy trung bình kết quả thực nghiệm.

Các phương pháp tư vấn cơ sở được thực nghiệm với các giá trị tham số khác nhau và lựa chọn kết quả độ chính xác tốt nhất để so sánh với các phương pháp đề xuất bởi luận án. Chi tiết về miền giá trị cho các tham số của các phương pháp lọc cộng tác và lọc theo nội dung cơ sở trình bày trong những công trình nghiên cứu liên quan được tham chiếu ở trên.

3.4.2.3. Kết quả kiểm nghiệm

Bảng 3.14. Giá trị MAE, RMSE của các phương pháp tư vấn trên MovieLens-1M

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**

Kết quả trong Bảng 3.14 cho thấy phương pháp tư vấn nội dung dựa vào hồ sơ người dùng và hồ sơ sản phẩm cho lại giá trị MAE, RMSE lớn nhất so với các phương pháp còn lại. Phương pháp tư vấn cộng tác dựa vào đánh giá người dùng và đánh giá sản phẩm cho lại giá trị MAE, RMSE nhỏ hơn so với các phương pháp tư vấn theo nội dung. Cụ thể, ứng với số lượng đánh giá biết trước trong tập kiểm tra là 5, 10, 15, 20, phương pháp *CBF-UerBased* và *CBF-Itembased* cho lại giá trị MAE lần lượt là 0.865, 0.859, 0.855, 0.835 và 0.894, 0.883, 0.876, 0.845 theo thứ

tự. Trong khi đó, phương pháp *CF-UserBased* và *CF-ItemBased* cho lại giá trị MAE lần lượt là 0.824, 0.817, 0.821, 0.813 và 0.846, 0.841, 0.836, 0.815 theo thứ tự. Nhận định tương tự khi so sánh giá trị RMSE của các phương pháp. Kết quả này hoàn toàn phù hợp với những nghiên cứu trước đây [10][47][115].

Phương pháp *Hybrid-UserBased* cho lại giá trị MAE thấp hơn nhiều so với phương pháp *CBF-UserBased* và *CF-UserBased*. Phương pháp *Hybrid-ItemBased* cũng cho lại giá trị MAE thấp hơn so với phương pháp *CBF-ItemBased* và *CF-ItemBased*. Điều này chỉ có thể lý giải phương pháp tính toán mức độ tương tự giữa các cặp người dùng dựa vào tập đánh giá người dùng cùng các đặc trưng sản phẩm chính xác hơn so với phương pháp tính toán mức độ tương tự giữa các cặp người dùng chỉ dựa vào đánh giá người dùng hoặc hồ sơ người dùng. Phương pháp tính toán mức độ tương tự giữa các cặp sản phẩm dựa vào tập đánh giá sản phẩm cùng các đặc trưng người dùng chính xác hơn so với phương pháp tính toán mức độ tương tự giữa các cặp sản phẩm chỉ dựa vào đánh giá sản phẩm hoặc hồ sơ sản phẩm.

Phương pháp *CoTraining- HybridFiltering* cho lại giá trị MAE, RMSE thấp nhất ở tất cả các mức độ thưa thớt dữ liệu khác nhau. Đặc biệt, với tập dữ liệu kiểm tra có 20 đánh giá biết trước, phương pháp cho lại giá trị MAE là 0.585 và RMSE là 0.707. 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.

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. Khi đó bài toán tư vấn kết hợp được dịch chuyển về bài toán tư vấn cộng tác. Sử dụng hợp nhất biểu diễn các giá trị đặc trưng nội dung vào lọc cộng tác để xây dựng phương pháp dự đoán cho lọc kết hợp bằng phương pháp đồng huấn luyện. 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 CHUNG

I. Kết quả đạt được của luận án

Luận án hướng tới một chủ đề có ý nghĩa về lý thuyết và thực tiễn của khoa học máy tính được cộng đồng nghiên cứu quan tâm, đó là nghiên cứu phát triển một số phương pháp xây dựng hệ tư vấn.

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. Những lý thuyết này về cơ bản dựa trên những kiến thức nền tảng của lý thuyết thống kê và học máy. 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; 2) Đề xuất một phương pháp lọc kết hợp bằng phương pháp đồng huấn luyện.

    - Với đề xuất chính thứ nhất, tác giả trình bày phương pháp hạn chế ảnh hưởng của vấn đề dữ liệu thưa của lọc cộng tác dựa trên mô hình đồ thị, mở rộng cho phát triển hệ tư vấn theo ngữ ảnh. Nội dung đề xuất này được trình bày chi tiết trong chương 2, kết quả đạt được được tổng hợp từ nghiên cứu đã công bố trong [C1][C3][C7][C4][J2] của tác giả.
    - Với đề xuất chính thứ hai, tác giả trình bày phương pháp kết hợp giữa lọc cộng tác và lọc nội dung bằng phương pháp đồng huấn luyện. Đây là một phương pháp thuộc hướng tiếp cận học bán giám sát cho bài toán phân lớp cho phép giải quyết vấn đề ít dữ liệu huấn luyện bằng cách chuyển giao kết quả huấn luyện qua lại khi quan sát theo người dùng và theo sản phẩm, nhờ vậy cải thiện độ chính xác dự đoán đánh giá và tư vấn sản phẩm phù hợp cho người dùng. Nội dung đề xuất này được trình bày chi tiết trong chương 3, kết

quả đạt được được tổng hợp từ nghiên cứu đã công bố trong [C2][C5][C6][J1] của tác giả.

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

Mặc dù các đề xuất được đưa ra bởi luận án giải quyết khá tốt vấn đề dữ liệu đánh giá thưa. Tuy nhiên vẫn còn tồn tại 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

Một số hướng nghiên cứu tác giả dự định thực hiện thời gian tới như sau:

- 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.

- Ngoài ra, các vấn đề chưa được giải quyết trong luận án như vấn đề người dùng mới, sở thích của người dùng với sản phẩm thay đổi theo thời gian cũng sẽ được tập trung nghiên cứu trong thời gian tới. Bên cạnh đó, tác giả cũng dự định nghiên cứu các phương pháp học dựa trên vùng quan tâm (Attention-based) để giải quyết bài toán lọc cộng tác và lọc kết hợp nhằm nâng cao chất lượng trong hệ tư vấn thực tế.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ

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

TÀI LIỆU THAM KHẢO

[1] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Knowledge-Based Systems Recommender systems survey,” *Knowledge-Based Syst.*, vol. 46, pp. 109–132, 2013.

[3] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and evaluating choices in a virtual community of use,” *Proc. SIGCHI Conf. Hum. factors Comput. Syst. - CHI ’95*, pp. 194–201, 1995.

[4] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: A survey,” *Decis. Support Syst.*, 2015.

[5] P. M. V.Sindhwani;, “Recommender systems,” *Commun. ACM*, pp. 1–21, 2010.

[6] G. Adomavicius, B. Mobasher, F. Ricci, and A. Tuzhilin, “Context-Aware Recommender Systems,” *AI Mag.*, vol. 32, no. 3, pp. 67–80, 2011.

[7] F. Ricci, L. Rokach, B. Shapira, and P. B.Kantor, *Recommender systems handbook*. Springer, 2011.

[8] J. J. P. A. F. V. P. D. Redondo, *Recommender Systems for the Social Web*. 2012.

[9] L. L. P. Mingsheng Fu, Hong Qu, Dagmawi Moges, “Attention Based Collaborative Filtering,” *Neurocomputing*, vol. 311, pp. 88–98, 2018.

[10] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Adv. Artif. Intell.*, vol. 2009, 2009.

[11] J. Xu, K. Johnson-wahrmann, and S. Li, “The Development , Status and Trends of Recommender Systems : A Comprehensive and Critical Literature Review,” *Math. Comput. Sci. Ind.*, vol. 7, no. 4, pp. 117–122, 2013.

[12] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems:

Principles, methods and evaluation," *Egypt. Informatics J.*, vol. 16, no. 3, pp. 261–273, 2015.

[13] B. Sarwar, "Item-Based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 285–295.

[14] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.

[15] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *UAI'98 Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 1998, pp. 43–52.

[16] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," *Expert Syst. Appl.*, vol. 97, pp. 205–227, 2018.

[17] B. Mobasher, X. Jin, and Y. Zhou, "Semantically enhanced collaborative filtering on the web," *Lect. Notes Comput. Sci.*, vol. 3209, no. 49, pp. 57–76, 2004.

[18] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," *Data MiningConcepts Tech.*, pp. 3–26, 2000.

[19] L. H. Ungar and D. P. Foster, "Clustering methods for collaborative filtering," *AAAI Work. Recomm. Syst.*, pp. 114–129, 1998.

[20] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," *Proc. 2000 ACM Conf. Comput. Support. Coop. Work - CSCW '00*, pp. 241–250, 2000.

[21] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, 2003.

[22] S. Gabrielsson, S. Gabrielsson, S. Gabrielsson, and S. Gabrielsson, "The use

of Self-Organizing Maps in Recommender Systems," *IEEE Trans. Neural Networks*, 2006.

[23] L. Ren and W. Wang, "An SVM-Based Collaborative Filtering Approach for Top-N Web Services Recommendation," *Futur. Gener. Comput. Syst.*, 2017.

[24] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proc. 23rd Int. Conf. Mach. Learn.*, vol. C, no. 1, pp. 161–168, 2006.

[25] C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: using social and content-based information in recommendation," in *AAAI '98/IAAI '98 Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998, pp. 714–720.

[26] X. Su, T. M. Khoshgoftaar, X. Zhu, and R. Greiner, "Imputation-boosted collaborative filtering using machine learning classifiers," *Proc. 2008 ACM Symp. Appl. Comput. - SAC '08*, no. 2, p. 949, 2008.

[27] D. Billsus and M. J. Pazzani, "User modeling for adaptive news access," *User Model. User-Adapted Interact.*, vol. 10, no. 2–3, pp. 147–180, 2000.

[28] X. Su and T. M. Khoshgoftaar, "Collaborative filtering for multi-class data using belief nets algorithms," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, pp. 497–504, 2006.

[29] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer (Long. Beach. Calif).*, vol. 42, no. 8, pp. 30–37, 2009.

[30] S. Zhang, L. Yao, and A. Sun, "Deep Learning based Recommender System: A Survey and New Perspectives," *CoRR, abs/1707.07435*, vol. 1, no. 1, pp. 1–35, 2017.

[31] M. Fu, H. Qu, and Z. Yi, "A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System," pp. 1–13, 2018.

[32] D. Jannach and M. Ludewig, "When Recurrent Neural Networks meet the

Neighborhood for Session-Based Recommendation," *Proc. Elev. ACM Conf. Recomm. Syst. - RecSys '17*, pp. 306–310, 2017.

[33] D. Cai, X. He, and J. W. W. Ma, "Block-level Link Analysis," in *SIGIR '04 Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 440–447.

[34] Z. Huang, W. Chung, and H. Chen, "A Graph Model for E-Commerce Recommender Systems," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 3, pp. 259–274, 2004.

[35] Z. Huang, H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 116–142, Jan. 2004.

[36] C. Filtering, R. Algorithms, I. N. E. Application, and A. Using, "A Study on Ontology Based Collaborative Filtering Recommendation Algorithms in E-Commerce Applications A Study on Ontology Based Collaborative Filtering Recommendation Algorithms in E-Commerce Applications," no. September, 2017.

[37] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Syst. Appl.*, vol. 92, pp. 507–520, 2018.

[38] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," 2016.

[39] C. A. Gomez-uribe and N. Hunt, "The Netflix Recommender System : Algorithms , Business Value ," vol. 6, no. 4, 2015.

[40] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," vol. 40, no. 3, 1997.

[41] M. J. Pazzani, "A Framework for Collaborative , Content-Based and Demographic Filtering," *Artif. Intell. Rev. - Spec. issue data Min. Internet*, vol. 13, no. 5–6, pp. 393–408, 1999.

[42] M. Pazzani and D. Billsus, "Learning and Revising User Profiles : The

Identification of Interesting Web Sites," *Mach. Learn. - Spec. issue multistrategy Learn.*, vol. 27, no. 3, pp. 313–331, 1997.

[43] G. L. Somlo, A. E. Howe, G. L. Somlo, and A. E. Howe, "Adaptive Lightweight Text Filtering Adaptive Lightweight Text Filtering," in *IDA 2001: Advances in Intelligent Data Analysis*, 2001, vol. 2189, pp. 319–329.

[44] Y. Zhang and J. Callan, "Maximum Likelihood Estimation for Filtering Thresholds," in *SIGIR '01 Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 294–302.

[45] R. J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," in *Proceedings of the SIGIR-99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, CA, August 1999*, 1999, pp. 195–204.

[46] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," *"Proceedings 18th Natl. Conf. Artif. Intell. (AAAI),"* no. July, pp. 187–192, 2002.

[47] R. Burke, "Hybrid Recommender Systems : Survey and Experiments," *User Model. User-adapt. Interact.*, vol. 12, no. 4, pp. 331–370, 2002.

[48] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," *ACM SIGIR Work. Recomm. Syst.*, 1999.

[49] D. Billsus and M. J. Pazzani, "A hybrid user model for news story classification," *UM99 User Model.*, pp. 99–108, 1999.

[50] B. Smyth and P. Cotter, "Personalized TV listings service for the digital TV age," *Knowledge-Based Syst.*, vol. 13, no. 2, pp. 53–59, 2000.

[51] A. M. Ahmad Wasfi, "Collecting user access patterns for building user profiles and collaborative filtering," *Proc. 4th Int. Conf. Intell. user interfaces - IUI '99*, pp. 57–64, 1999.

[52] R. D. Burke, K. J. Hammond, and B. C. Young, "The FindMe approach to

assisted browsing," *IEEE Expert. Syst. their Appl.*, vol. 12, no. 4, pp. 32–40, 1997.

[53] N. Good, J. Ben Schafer, J. A. Konstan, A. Borchers, and B. Sarwar, "Combining Collaborative Filtering with Personal Agents for Better Recommendations Nathaniel," *Tetrahedron*, vol. 62, no. 37, pp. 8805–8813, 2006.

[54] N. D. Phuong, L. Q. Thang, and T. M. Phuong, "A Graph-Based Method for Combining Collaborative and Content-Based Filtering," in *Proceedings of PRICAI 2008*, 2008, vol. 5351, pp. 859–869.

[55] A. Gunawardana and C. Meek, "A unified approach to building hybrid recommender systems," *Proc. third ACM Conf. Recomm. Syst. - RecSys '09*, p. 117, 2009.

[56] A. P. and L. H.Ungar and D. M. P. and S. Lawrence, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," in *UAI'01 Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 2001, pp. 437–444.

[57] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '02*, no. August, p. 253, 2002.

[58] A. Ansari, S. Essegaier, and R. Kohli, "Internet Recommendation Systems," *J. Mark. Res.*, vol. 37, no. 3, pp. 363–375, 2000.

[59] A. V. I. Arampatzis and G. Kalamatianos, "Collaborative , and Hybrid Fusion Methods," vol. 36, no. 3, 2017.

[60] R. Xiong, J. Wang, N. Zhang, and Y. Ma, "Deep hybrid collaborative filtering for Web service recommendation," *Expert Syst. Appl.*, vol. 110, pp. 191–205, 2018.

[61] F. Ortega, D. Rojo, and L. Raya, "Hybrid Collaborative Filtering based on Users Rating Behavior," vol. XX, no. c, 2018.

[62] T. Xiao, S. Liang, H. Shen, and Z. Meng, "Neural Variational Hybrid

Collaborative Filtering," 2019.

[63] Y. Tay, L. A. Tuan, and S. C. Hui, "Latent Relational Metric Learning via Memory-based Attention for Collaborative Ranking," 2018.

[64] L. Zheng, C. Lu, V. Noroozi, H. Huang, and P. S. Yu, "MARS : Memory Attention-Aware Recommender System," no. July 2017, 2018.

[65] Z. Cheng, Y. Ding, X. He, L. Zhu, X. Song, and M. Kankanhalli, "A 3 NCF : An Adaptive Aspect Attention Model for Rating Prediction," pp. 3748–3754, 2009.

[66] M. Hahsler, "recommenderlab : A Framework for Developing and Testing Recommendation Algorithms." 2011.

[67] L. Si and R. Jin, "Flexible Mixture Model for Collaborative Filtering," *Mach. Learn. Work.*, vol. 20, no. 2, p. 704, 2003.

[68] D. Billsus and M. J. Pazzani, "Learning Collaborative Information Filters," in *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 46–54.

[69] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste : A Constant Time Collaborative Filtering Algorithm," no. August, 2000.

[70] B. M. Sarwar, G. Karypis, J. a Konstan, and J. T. Riedl, "Application of Dimensionality Reduction in Recommender System - A Case Study," *Architecture*, vol. 1625, pp. 264–8, 2000.

[71] N. Srebro and T. Jaakkola, "Weighted Low-Rank Approximations," 2003.

[72] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-Walk Computation of Similarities Between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[73] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decis. Support Syst.*, vol. 54, no. 2, pp. 880–890, 2013.

[74] K. Yang and L. Toni, "Graph-Based Recommendation System," 2018.

[75] L. Deladiennee and Y. Naudet, “A graph-based semantic recommender system for a reflective and personalised museum visit: Extended abstract,” in *Proceedings - 12th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2017*, 2017, pp. 88–89.

[76] U. Panniello, A. Tuzhilin, and M. Gorgoglione, “Comparing context-aware recommender systems in terms of accuracy and diversity,” *User Model. User-adapt. Interact.*, vol. 24, no. 1–2, pp. 35–65, 2014.

[77] L. Baltrunas, B. Ludwig, and F. Ricci, “Matrix Factorization Techniques for Context Aware,” *Acm Rs*, no. October, pp. 301–304, 2011.

[78] A. Q. Macedo, C. Grande, and C. Grande, “Context-Aware Event Recommendation in Event-based Social Networks Categories and Subject Descriptors,” *2015 ACM Conf. Recomm. Syst. RecSys 2015*, pp. 123–130, 2015.

[79] N. X. Bach, N. Do Hai, and T. M. Phuong, “Personalized recommendation of stories for commenting in forum-based social media ☆,” *Inf. Sci. (Ny).*, vol. 352–353, pp. 48–60, 2016.

[80] H. Yin and B. Cui, *Spatio-Temporal Recommendation in Social Media*. 2016.

[81] L. Cai, J. Xu, J. Liu, and T. Pei, “Integrating spatial and temporal contexts into a factorization model for POI recommendation,” *Int. J. Geogr. Inf. Sci.*, vol. 32, no. 3, pp. 524–546, 2018.

[82] A. Razia Sulthana and S. Ramasamy, “Ontology and context based recommendation system using Neuro-Fuzzy Classification,” *Comput. Electr. Eng.*, vol. 0, pp. 1–13, 2018.

[83] X. Fan, Y. Hu, Z. Zheng, Y. Wang, P. Brezillon, and W. Chen, “CASR-TSE: Context-Aware Web Services Recommendation for Modeling Weighted Temporal-Spatial Effectiveness,” *IEEE Trans. Serv. Comput.*, p. 1, 2017.

[84] M. Afzal, S.I. Ali, R. Ali, M. Hussain, T. Ali, W.A. Khan, M.B. Amin, B.H. Kang, S. Lee, “Personalization of wellness recommendations using contextual interpretation,” *Expert Syst. Appl.*, vol. 96, pp. 506–521, 2018.

[85] S. L. Wang and C. Y. Wu, “Application of context-aware and personalized recommendation to implement an adaptive ubiquitous learning system,” *Expert Syst. Appl.*, vol. 38, no. 9, pp. 10831–10838, 2011.

[86] K. Tang, S. Chen, and A. J. Khattak, “Personalized travel time estimation for urban road networks: A tensor-based context-aware approach,” *Expert Syst. Appl.*, vol. 103, pp. 118–132, 2018.

[87] A. Garcia-de-Prado, G. Ortiz, and J. Boubeta-Puig, “COLLECT: COLLaborativE ConText-aware service oriented architecture for intelligent decision-making in the Internet of Things,” *Expert Syst. Appl.*, vol. 85, pp. 231–248, 2017.

[88] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, “Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach,” *ACM Trans. Inf. Syst.*, vol. 23, no. 1, pp. 103–145, Jan. 2005.

[89] L. Baltrunas and F. Ricci, “Experimental evaluation of context-dependent collaborative filtering using item splitting,” *User Model. User-adapt. Interact.*, vol. 24, no. 1–2, pp. 7–34, 2013.

[90] Y. Zheng, R. Burke, and B. Mobasher, “Differential Context Relaxation for Context-Aware Travel Recommendation,” in *International Conference on Electronic Commerce and Web Technologies.EC-Web 2012*, 2012, vol. 123, no. September, pp. 88–99.

[91] V. Codina, F. Ricci, and L. Ceccaroni, “Exploiting the semantic similarity of contextual situations for pre-filtering recommendation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7899 LNCS, pp. 165–177, 2013.

[92] B. Zou, C. Li, L. Tan, and H. Chen, “GPUTENSOR: Efficient tensor factorization for context-aware recommendations,” *Inf. Sci. (Ny).*, vol. 299, pp. 159–177, 2015.

[93] Y. Zheng, B. Mobasher, and R. Burke, “CSLIM:Contextual SLIM

Recommendation Algorithms," *RecSys 2014 - Proc. 8th ACM Conf. Recomm. Syst.*, vol. 0, no. 1, pp. 301–304, 2014.

[94] Y. Zheng, B. Mobasher, and R. Burke, "Deviation-Based Contextual SLIM Recommenders," *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '14*, no. Dcm, pp. 271–280, 2014.

[95] K. Haruna, M.A. Ismail, S. Suhendroyono, D. Damiasih, A.C. Pierewan, H. Chiroma, T. Herawan, "Context-Aware Recommender System: A Review of Recent Developmental Process and Future Research Direction," *Appl. Sci.*, vol. 7, no. 12, p. 1211, 2017.

[96] S. Lee, S. Song, M. Kahng, D. Lee, and S. Lee, "Random walk based entity ranking on graph for multidimensional recommendation," *Proc. 5th ACM Conf. Recomm. Syst. - RecSys '11*, p. 93, 2011.

[97] A. R. Ana, Á. M. G. Carvalho, and C. G. Ralha, "Agent-based architecture for context-aware and personalized event recommendation," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 563–573, 2014.

[98] P. Bedi and Richa, "User interest expansion using spreading activation for generating recommendations," *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 766–771, 2015.

[99] E. Şamdan and A. Taşcı, "A Graph-based Collaborative and Context-aware Recommendation system for TV programs," in *RecSys 2014 TV Workshop*, 2014, no. October, pp. 1–6.

[100] Z. Bahramian, R. A. Abbaspour, and C. Claramunt, "A context-aware tourism recommender system based on a spreading activation method," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 42, no. 4W4, pp. 333–339, 2017.

[101] L. Baltrunas and F. Ricci, "Context-Based Splitting of Item Ratings in Collaborative Filtering," in *Proceedings of the third ACM conference on Recommender systems - RecSys '09*, 2009, pp. 245–248.

[102] Y. Zheng, R. Burke, and B. Mobasher, "Splitting approaches for context-

aware recommendation," *Proc. 29th Annu. ACM Symp. Appl. Comput. - SAC '14*, pp. 274–279, 2014.

[103] Y. Zheng, B. Mobasher, and R. Burke, "CARSKit: A Java-Based Context-Aware Recommendation Engine," in *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1668–1671.

[104] L. Baltrunas and F. Ricci, "Experimental evaluation of context-dependent collaborative filtering using item splitting," *User Model. User-adapt. Interact.*, vol. 24, no. 1–2, pp. 7–34, 2014.

[105] X. Ning and G. Karypis, "SLIM: Sparse Linear Methods for Top-N Recommender Systems," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, 2011, pp. 497–506.

[106] Y. Zheng, "Tutorial : Context In Recommender Systems," in *The 31st ACM Symposium on Applied Computing*, 2016.

[107] N. D. Phuong and T. M. Phuong, "Collaborative Filtering by Multi-task Learning," vol. 00, no. c, pp. 1–6, 2008.

[108] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 2014.

[109] A. Z. Olivier Chapelle, Bernhard Scho ̈lkopf, *A semi-supervised learning*, vol. 1, no. 2. The MIT Press Cambridge, Massachusetts London, England, 2009.

[110] P. Rai, "Semi-supervised Learning," in *CS 5350/6350: Machine Learning*, 2011, vol. 2011.

[111] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proc. Elev. Annu. Conf. Comput. Learn. theory - COLT' 98*, pp. 92–100, 1998.

[112] C. Guestrin, "Co-Training for Semi- supervised learning," in *Machine Learning - 10701/15781*, 2007, pp. 1–51.

[113] W. Wang and Z.-H. Zhou, "A New Analysis of Co-Training," *ICML Int. Conf. Mach. Learn.*, pp. 1135–1142, 2011.

[114] A. Gunawardana, “A Survey of Accuracy Evaluation Metrics of Recommendation Tasks,” *J. Mach. Learn. Res. 10*, vol. 10, pp. 2935–2962, 2009.

[115] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen, “Collaborative Filtering Recommender Systems,” *Adapt. Web Methods Strateg. Web Pers.*, vol. 4321, pp. 291–324, 2007.