MỘT PHƯƠNG PHÁP LỌC TRƯỚC THEO NGỮ CẢNH CHO HỆ TƯ VẤN

Đỗ Thị Liên

Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Hệ tư vấn là 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. Bên cạnh những thông tin phổ biến (người dùng, sản phẩm, đánh giá của người dùng với sản phẩm) được khai thác thường xuyên trong hệ tư vấn, một trong những yếu tố quan trọng ảnh hưởng tới việc ra quyết định trong 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. Mặc dù việc tích hợp ngữ cảnh vào hệ tư vấn được chứng minh là giúp nâng cao đáng kể chất lượng tư vấn sản phẩm tới người dùng, tuy nhiên khó khăn điển hình gặp phải đối với hệ tư vấn dựa vào ngữ cảnh lúc này là làm thế nào để tích hợp thông tin ngữ cảnh và vấn đề dữ liệu thưa, điều này ảnh hưởng trực tiếp tới chất lượng tư vấn. Trong bài báo này, tác giả đề xuất một phương pháp lọc trước theo ngữ cảnh cho hệ tư vấn cho phép tích hợp đầy đủ thông tin ngữ cảnh và giải quyết hiệu quả vấn đề dữ liệu thưa. Trong đó, việc tích hợp ngữ cảnh được thực hiện bằng thủ tục phân tách sản phẩm theo ngữ cảnh và vấn đề dữ liệu thưa được giải quyết qua quá trình huấn luyện theo mô hình đồng huấn luyện cho bài toán phân lớp của lọc cộng tác. Kết quả thực nghiệm trên một số bộ dữ liệu thực cho thấy phương pháp đề xuất cải thiện đáng kể chất lượng dự đoán so với các phương pháp tư vấn dựa vào ngữ cảnh cơ sở trước đây.

Từ khóa: Hệ tư vấn dựa vào ngữ cảnh (Context-aware recommender system - CARS); Lọc cộng tác dựa vào ngữ cảnh (Context-aware collaborative filtering - CACF); Ngữ cảnh (Context); Lọc trước theo ngữ cảnh (Contextual pre-filtering); Phân tách sản phẩm theo ngữ cảnh (Item splitting); Mô hình đồng huấn luyện (Co-training model).

I. MỞ ĐẦU

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 sản phẩm phù hợp với sở thích riêng của từng người. Các nghiên cứu về việc xây dựng hệ tư vấn truyền thống trước đây chủ yếu tập trung sử dụng thông tin của người dùng, sản phẩm và đánh giá của người dùng với sản phẩm trong việc đưa ra gợi ý. Bài toán tư vấn truyền thống có thể được biểu diễn dựa trên ma trận đánh giá hai chiều sau:

$$R_o: User \times Item \rightarrow Rating \quad (1)$$

Để giải quyết bài toán tư vấn truyền thống, các nghiên cứu đã có được tiếp cận theo ba hướng chính là: lọc cộng tác, lọc theo nội dung và lọc kết hợp [1].

Tuy nhiên trên thực tế, đánh giá của người dùng không cố định, mà thay đổi theo ngữ cảnh bên ngoài. Ví dụ một người trời nóng thì thích ăn kem, nhưng trời lạnh lại thích ăn lẩu. Hoặc cùng một bộ phim nhưng trời mưa thì thích xem còn trời khô ráo thì có khi lại không thích. Do vậy việc xem xét kết hợp thông tin ngữ cảnh vào các hệ tư vấn đang là chủ đề rất được quan tâm nghiên cứu trong những năm gần đây. Theo như [2]: "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". 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). 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. Ví dụ một số hệ tư vấn dựa trên ngữ cảnh như hệ tư vấn phim và âm nhạc [3], tư vấn địa điểm yêu thích [4], thương mại điện tử [5],…Khi đó, bài toán tư vấn theo ngữ cảnh sẽ được biểu diễn dựa trên ma trận đánh giá đa chiều (Multi-dimensional matrix) như sau:

$$R_1: User \times Item \times Context \rightarrow Rating \quad (2)$$

Để giải quyết bài toán tư vấn theo ngữ cảnh, có 3 hướng tiếp cận điển hình được biết đến là: Lọc trước theo ngữ cảnh (Contextual Prefiltering), lọc sau theo ngữ cảnh (Contextual Postfiltering) và mô hình hóa ngữ cảnh (Contextual Modeling) [2][6]. Trong đó, lọc trước ngữ cảnh sử dụng thông tin ngữ cảnh để lọc tập dữ liệu ban đầu nhằm chỉ giữ lại những dữ liệu phù hợp với ngữ cảnh yêu cầu, dữ liệu lọc được sẽ được học bởi các phương pháp tư vấn truyền thống. Trái ngược với hướng lọc trước theo ngữ cảnh, lọc sau theo ngữ cảnh sẽ sử dụng các phương pháp tư vấn truyền thống để học dữ liệu, kết quả dự đoán thu được sẽ được lọc lại một lần nữa bởi ngữ cảnh hiện thời nhằm thu được kết quả tư vấn cuối cùng. Đối với hai hướng tiếp cận lọc trước theo ngữ cảnh và lọc sau theo ngữ cảnh thì thông tin ngữ cảnh không được tích hợp trong quá trình huấn luyện. Một hướng tiếp cận thứ ba cho phép tích hợp trực tiếp ngữ cảnh vào quá trình huấn luyện và tư vấn đó là mô hình hóa ngữ cảnh. Theo hướng tiếp cận thứ ba này thì 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. Căn cứ vào kết quả thực nghiệm nhiều nghiên cứu đã chỉ ra rằng mỗi phương pháp đều có những ưu nhược điểm riêng, không có phương pháp nào là tốt cho mọi trường hợp dữ liệu, việc lựa chọn phương pháp nào sẽ phụ thuộc vào hiệu quả cho từng bộ dữ liệu của bài toán nghiệp vụ khác nhau [7], 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

Tác giả liên hệ: Đỗ Thị Liên
Email: liendt@ptit.edu.vn
Đến tòa soạn: 10/2020, chỉnh sửa: 11/2020 , chấp nhận đăng: 12/2020

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.

Nhằm phát huy tính đơn giản trong cài đặt và tận dụng được các phương pháp tư vấn truyền thống đã có, trong bài báo này tác giả tiếp cận các phương pháp lọc trước theo ngữ cảnh là đối tượng nghiên cứu của mình. Theo hướng này, có một số nghiên cứu điển hình như: phương pháp phân tách sản phẩm theo ngữ cảnh (Item splitting) được đề xuất bởi Baltrunas và Ricci[8], phương pháp phân cụm người dùng và sản phẩm dựa trên điều kiện ngữ cảnh [9], phương pháp thu giảm số chiều nhằm giữ lại trong tập giữ liệu một tập nhỏ hơn những ngữ cảnh quan trọng [10]. Theo đó, phương pháp phân tách sản phẩm theo ngữ cảnh chỉ sử dụng một chiều ngữ cảnh để phân tách sản phẩm ban đầu thành các sản phẩm giả lập, khiến có khá nhiều thông tin từ các chiều ngữ cảnh khác không được khai thác triệt để vào quá trình tư vấn sau này. Phương pháp phân cụm theo ngữ cảnh cũng gặp vấn đề cần thay đổi theo mỗi tập dữ liệu khác nhau, dẫn tới khó khăn trong việc kết hợp kết quả tư vấn sau đó. Phương pháp thu giảm số chiều cũng gặp vấn đề có thể bỏ qua thông tin hữu ích từ những chiều ngữ cảnh khác phục vụ cho quá trình huấn luyện dữ liệu.

Để giảm thiểu những hạn chế nêu trên, tác giả đề xuất một phương pháp tư vấn cộng tác theo ngữ cảnh mới thuộc hướng tiếp cận lọc trước ngữ cảnh nhằm giải quyết hạn chế còn tồn tại phổ biến, đó là tích hợp đầy đủ thông tin ngữ cảnh và giải quyết hiệu quả vấn đề dữ liệu thưa của hệ tư vấn theo ngữ cảnh. Cụ thể, tác giả đề xuất sử dụng 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, đó là cho phép tích hợp đầy đủ thông tin ngữ cảnh trong việc chuyển hóa sản phẩm ban đầu thành sản phẩm giả lập. Tuy nhiên khi áp dụng thủ tục phân tách sản phẩm theo ngữ cảnh cải tiến lên ma trận đánh giá đa chiều R_1, với việc giới thiệu các sản phẩm giả lập, sẽ càng khiến ma trận đánh giá hai chiều R_o thu được càng thưa thớt hơn nữa, điều này sẽ được giải quyết qua quá trình học dữ liệu đánh giá theo mô hình đồng huấn luyện cho bài toán phân lớp của lọc cộng tác. Đồng huấn luyện là một phương pháp học bán giám sát điển hình sử dụng để học từ cả những đánh giá đã biết và chưa biết trong ma trận đánh giá để đưa ra dự đoán, từ đó giải quyết hiệu quả vấn đề dữ liệu thưa của ma trận đánh giá, giúp nâng cao chất lượng tư vấn. Kết quả thực nghiệm trên một số bộ dữ liệu thực cho thấy phương pháp đề xuất cải thiện đáng kể chất lượng dự đoán so với các phương pháp tư vấn dựa vào ngữ cảnh cơ sở trước đây. Phương pháp đề xuất cũng được đánh giá là đơn giản trong cài đặt và tận dụng được các phương pháp tư vấn truyền thống đã có.

Để trọng tâm vào phương pháp đề xuất, Mục II tác giả trình bày các nghiên cứu liên quan. Tiếp đến là phương pháp đề xuất trong Mục III. Mục IV trình bày phương pháp thực nghiệm và đánh giá. Mục V nêu kết luận và hướng phát triển trong thời gian tới.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Việc khai thác thông tin ngữ cảnh trong hệ tư vấn được đề cập đầu tiên trong nghiên cứu của Adomavicius and Tuzhilin [1], theo đó dữ liệu huấn luyện được biểu diễn từ sự kết hợp của người dùng, sản phẩm và các chiều ngữ cảnh tương ứng. Trên cơ sở đó, có khá nhiều công trình nghiên cứu liên quan khác được công bố sau đó nhằm giải quyết các vấn đề khác nhau của hệ tư vấn theo ngữ cảnh, trong đó tập trung vào cải tiến các phương pháp tư vấn theo ngữ cảnh.

Các phương pháp đưa ra để giải quyết bài toán tư vấn theo ngữ cảnh thuộc ba hướng tiếp cận: (i) lọc trước ngữ cảnh, (ii) lọc sau ngữ cảnh và (iii) mô hình hóa ngữ cảnh [2][6].

Về cơ bản, các phương pháp tư vấn theo ngữ cảnh thuộc hướng lọc trước ngữ cảnh sử dụng thông tin ngữ cảnh để lọc tập dữ liệu ban đầu nhằm chỉ giữ lại những dữ liệu phù hợp với ngữ cảnh yêu cầu.Trong đó, phân tách sản phẩm theo ngữ cảnh (Item splitting) [8] là một phương pháp điển hình thuộc hướng tiếp cận này được đánh giá cho hiệu quả tư vấn tương đối tốt trên nhiều tập dữ liệu. Phương pháp này thực hiện tách mỗi sản phẩm trong tập dữ liệu ban đầu thành các sản phẩm giả lập. Trong đó, mỗi sản phẩm giả lập được tạo ra từ sự kết hợp sản phẩm ban đầu với một tình huống ngữ cảnh cụ thể. Tập dữ liệu lọc được sẽ dùng để huấn luyện và tư vấn. Quá trình huấn luyện và tư vấn cho hệ tư vấn theo ngữ cảnh sau đó có thể sử dụng trực tiếp những phương pháp lọc thông tin đã được áp dụng cho các hệ tư vấn truyền thống. Ví dụ như một số phương pháp lọc cộng tác như UserKNN, ItemKNN, Matrix Factorization, SLIM…[2][6] sẽ được áp dụng trực tiếp sau bước lọc trước ngữ cảnh để sinh những sản phẩm dự đoán cho người dùng trong một tình huống ngữ cảnh cụ thể. Tương tự như thế, phân tách người dùng theo ngữ cảnh (User splitting), phân tách cả người dùng và sản phẩm theo ngữ cảnh (UI splitting) là các phương pháp được đưa ra thuộc hướng này có cơ chế hoạt động tương tự [11].

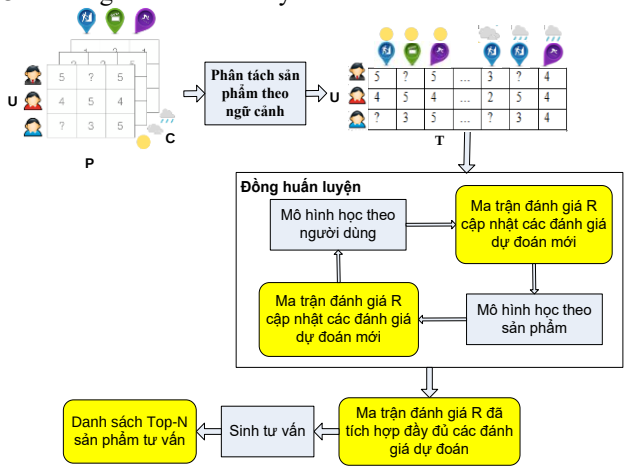
Trái ngược với hướng lọc trước ngữ cảnh, lọc sau ngữ cảnh sử dụng toàn bộ ma trận đánh giá đã loại bỏ đi các chiều ngữ cảnh để huấn luyện và tư vấn. Kết quả tư vấn sẽ được lọc lại một lần nữa để thu được kết quả tư vấn cuối cùng là những sản phẩm mới chưa được người dùng đánh giá trong một tình huống ngữ cảnh cụ thể. Như vậy các phương pháp tư vấn theo ngữ cảnh thuộc hướng lọc sau ngữ cảnh cũng có thể áp dụng các phương pháp tư vấn truyền thống như (i).

Hướng tiếp cận thứ ba là mô hình hóa ngữ cảnh. Theo hướng này 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 phương pháp mô hình hóa ngữ cảnh được đưa ra như: Mô hình hóa ngữ cảnh độc lập Tensor Decomposition [12] và mô hình hóa ngữ cảnh phụ thuộc [13][14]. 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 [14]. Tuy nhiên vấn đề đặt ra với các phương pháp mô hình hóa ngữ cảnh phụ thuộc là khi tích hợp ngữ cảnh vào hệ tư vấn dựa trên các giải thuật tư vấn truyền thống như Matrix Factorization, SLIM…là vấn đề dữ liệu thưa và khả năng mở rộng của nó. Ngoài ra các phương pháp mô hình hóa ngữ cảnh cũng được đánh giá là có độ phức tạp lớn hơn các phương pháp thuộc hướng lọc trước và sau theo ngữ cảnh, đặc biệt khi số chiều ngữ cảnh tăng lên.

Trong bài báo này, tác giả tiếp cận nghiên cứu đề xuất phương pháp tư vấn theo ngữ cảnh mới, cụ thể là một phương pháp tư vấn thuộc hướng lọc trước theo ngữ cảnh. Trong đề xuất của mình, tác giả tiếp cận kết hợp thủ tục phân tách sản phẩm theo ngữ cảnh cải tiến với mô hình đồng huấn luyện cho lọc cộng tác. Khác với việc sử dụng một chiều ngữ cảnh để phân tách sản phẩm ban đầu thành các sản phẩm giả lập không cho phép khai thác triệt để thông tin từ các chiều ngữ cảnh còn lại, thủ tục phân tách sản phẩm theo ngữ cảnh cải tiến sẽ kết hợp sử dụng tất cả các chiều ngữ cảnh để phân tách sản phẩm ban đầu thành các sản phẩm giả lập, điều này giúp tích hợp đầy đủ thông tin ngữ cảnh vào quá trình tư vấn. Việc giới thiệu các sản phẩm giả lập cũng đồng thời chuyển hóa ma trận đánh giá đa chiều thành ma trận đánh giá hai chiều. Tuy nhiên, bản thân ma trận đánh giá đa chiều ban đầu khá thưa khi được chuyển hóa thành ma trận đánh giá hai chiều càng trở lên thưa thớt hơn nữa. Để giải quyết vấn đề này, tác giả tiếp cận học dữ liệu đánh giá cho lọc cộng tác bằng mô hình đồng huấn luyện, đây là một phương pháp điển hình thuộc hướng học bán giám sát cho bài toán phân lớp, nhằm học từ cả những đánh giá đã biết và chưa biết trong ma trận đánh giá, từ đó giải quyết hiệu quả vấn đề dữ liệu thưa của ma trận đánh giá, giúp nâng cao chất lượng tư vấn cho hệ tư vấn theo ngữ cảnh.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Về cơ bản phương pháp đề xuất được thực hiện bằng cách kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh; 2) Lọc cộng tác bằng đồng huấn luyện. Sự kết hợp của hai phương pháp này trong phương pháp đề xuất được thể hiện qua ba bước: 1) Phân tách sản phẩm theo ngữ cảnh; 2) Học dữ liệu bằng mô hình đồng huấn luyện; 3) Sinh tư vấn. Ba bước này kết hợp với nhau trong một bộ khung đề xuất về triển khai phương pháp lọc trước theo ngữ cảnh dựa vào đồng huấn luyện cho hệ tư vấn theo ngữ cảnh trong Hình 1 dưới đây.



Hình 1. Bộ khung triển khai phương pháp lọc trước theo ngữ cảnh dựa vào đồng huấn luyện.

A. Phân tách sản phẩm theo ngữ cảnh

Thông tin đầu vào cho bài toán tư vấn theo ngữ cảnh 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" [8].

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) [3][11] . 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, tác giả 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 1) ta thu được ma trận đánh giá hai chiều (Bảng 2), 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 1, 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, tác giả 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.

Bảng 1. *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	Tại rạp	Đối tác

			tuần		
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

Bảng 2. *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.

Để hạn chế những vấn đề dữ liệu thưa của lọc cộng tác áp dụng cho ma trận đánh giá hai chiều R_o, tác giả tiến hành học ma trận đánh giá của lọc cộng tác bằng mô hình đồng huấn luyện. Nội dung chi tiết của mô hình đồng huấn luyện cho lọc cộng tác được trình bày trong mục B dưới đây.

B. Mô hình đồng huấn luyện cho lọc cộng tác

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 [15][16][17][18]. 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. 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.

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 [19][20], 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 bốn hướng tiếp cận học máy này thì 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. Với thông tin đầu vào của lọc cộng tác là ma trận đánh giá 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.

Trong các phương pháp học bán giám sát đã được đưa ra [21][22], tác giả tiếp cận phương pháp đồng huấn luyện để giải quyết bài toán phân lớp của lọc cộng tác. Lý do cho việc lựa chọn này là 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, trong trường hợp này mỗi đánh giá trong ma trận đánh giá sẽ được quan sát dưới hai góc nhìn người dùng và sản phẩm.

Quá trình đồng huấn luyện sẽ sử dụng 2 bộ phân lớp xác định nhằm học các mẫu dữ liệu độc lập từ quan sát theo người dùng và quan sát theo sản phẩm để gán nhãn cho các mẫu dữ liệu chưa biết, trong trường hợp này là đưa ra dự đoán đánh giá cho những giá trị đánh giá chưa biết trong ma trận đánh giá. 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 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 đánh giá. 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.

Nội dung mục 1) và 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 ma trận đánh giá R_o nhận được theo thủ tục phân tách sản phẩm theo ngữ cảnh (Mục A). Trên cơ sở đó đề xuất kết hợp hai mô hình này trong hai phương pháp đồng huấn luyện cho lọc cộng tác trong mục 3). Kết quả của quá trình đồng huấn luyện là ma trận R_o 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

1) Mô hình học theo người dùng

Mô hình học theo người dùng được sử dụng trong bài báo dựa vào phương pháp lọc cộng tác theo người dùng UserBased k-NN [1][23]. Đây là phương pháp được đánh giá là đơn giản trong cài đặt, thời gian thực hiện nhanh và có thể thực hiện được trên mọi loại dữ liệu. Tuy nhiên nhược điểm điển hình với phương pháp này là vấn đề dữ liệu thưa, điều này sẽ được tác giả giải quyết trong mô hình đồng huấn luyện đưa ra bởi bài báo.

Thuật toán UserBased k-NN đã có xây dựng cho mỗi người dùng một tập các láng giềng có các đánh giá tương tự trong ma trận người dùng – sản phẩm, các đánh giá từ

những người dùng này sau đó được sử dụng để đưa ra dự đoán, làm cơ sở để đưa ra tư vấn. Trong đề xuất này, việc xác định mức độ tương tự giữa các cặp người dùng không dùng để xác định tập láng giềng K_i tác động trực tiếp lên tư vấn, mà chỉ để dùng vào việc xác định các nhãn phân loại chắc chắn r_{iy} cho người dùng u_i. Theo đó, mô hình học theo người dùng đề xuất được thực hiện thông qua 3 bước: (1) Tính toán độ tương tự giữa các cặp người dùng; (2) Tìm tập láng giềng cho người dùng cần tư vấn; (3) Xác định các nhãn phân loại chắc chắn cho ma trận đánh giá.

Bước 1. Tính toán mức độ tương tự giữa các cặp người dùng

Độ tương tự dựa trên độ đo tương quan Pearson

$$u_{ij} = \frac{\sum_{p_x \in T_i \cap T_j}(r_{ix} - \overline{r_i})(r_{jx} - \overline{r_j})}{\sqrt{\sum_{t_x \in T_i \cap T_j}(r_{ix} - \overline{r_i})^2 \sum_{t_x \in T_i \cap T_j}(r_{jx} - \overline{r_j})^2}} \quad (3)$$

Trong đó :

- $T_{ij} = \{t_x | r_{ix} \neq \emptyset \cap r_{jx} \neq \emptyset\}$ là tập tất cả các sản phẩm cùng được đánh giá bởi u_i và u_j.
- $\overline{r_i}, \overline{r_j}$ là trung bình cộng các đánh giá khác 0 của u_i và u_j.

Bước 2: Xác định tập láng giềng cho người dùng cần tư vấn

Tại bước này ta chỉ cần sắp xếp các giá trị u_{ij} theo thứ tự giảm dần, sau đó chọn tập K_i người dùng đầu tiên làm tập láng giềng của người dùng u_i.

Bước 3: Xác định các nhãn phân loại chắc chắn cho ma trận

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 (4).

$$r_{ix} = \overline{r_i} + \frac{\sum_{u_j \in K_i}(r_{jx} - \overline{r_j})u_{ij}}{\sum_{u_j \in K_i}|u_{ij}|} \quad (4)$$

Ma trận đánh giá nhận được sau mô hình học theo người dùng được bổ sung các đánh giá dự đoán mới, phục vụ cho quá trình huấn luyện theo mô hình học theo sản phẩm trong phương pháp đồng huấn luyện được trình bày ở mục tiếp theo của bài báo.

2) Mô hình học theo sản phẩm

Mô hình học theo sản phẩm được sử dụng trong bài báo dựa vào phương pháp lọc cộng tác theo sản phẩm ItemBased k-NN [1][23].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][23], mà chỉ để dùng vào việc xác định các nhãn phân loại chắc chắn r_{ix} cho sản phẩm t_x. Ma trận đánh giá nhận được sau mô hình học theo sản phẩm được bổ sung các đánh giá dự đoán mới, phục vụ cho quá trình huấn luyện theo mô hình học theo người dùng trong phương pháp đồng huấn luyện được trình bày ở mục tiếp theo của bài báo.

3) Lọc cộng tác bằng phương pháp đồng huấn luyện

Phương pháp CoTrainning-UserItem được mô tả chi tiết trong Thuật toán 3.2 thực hiện thông qua t vòng lặp.

Thuật toán 1. Thuật toán CoTrainning-UserItem

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

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

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 mô hình học theo người dùng:

a) *Tìm* $u_{ij}^{(t)}$ *theo công thức* (3).

b) *Tìm* $K_i^{(t)}$

c) *Dự đoán* $r_{ix}^{(t)}$ *theo công thức* (4).

2.3. Huấn luyện theo mô hình học theo sản phẩm:

a) *Tìm* $t_{xy}^{(t)}$ căn cứ theo thuật toán ItemBased k-NN cơ sở [1][23].

b) *Tìm* $K_x^{(t)}$.

c) *Dự đoán* $r_{ix}^{(t)}$.

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

Tại bước khởi tạo $t = 0$, ma trận dự đoán $R_1{}^{(0)} = \{r_{ix}^{(0)}\}$ được lấy bằng chính ma trận đánh giá ban đầu $R_1 = \{r_{ix}\}$. Tại đầu mỗi bước lặp, t được tăng lên 1 đơn vị. Tại bước 2.2, quá trình huấn luyện theo người dùng được thực hiện tuần tự theo các bước (2.2.a), (2.2.b),(2.2.c). Tại bước (2.2.a) ta cần xác định mức độ tương tự $u_{ij}^{(t)}$ giữa người $u_i \in U$ và người dùng u_j theo công thức (3). Tại bước (2.2.b), sử dụng $u_{ij}^{(t)}$ đã được xác định tại bước (2.2.a) ta xác định được $K_i^{(t)}$ là tập láng giềng của người dùng u_i tại bước lặp thứ t. Tại bước (2.2.c), sử dụng $K_i^{(t)}$ 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 (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 2.3.

Tại bước (2.3.a) ta cần xác định mức độ tương tự $t_{xy}^{(t)}$ giữa sản phẩm $t_x \in T$ và sản phẩm t_y. Tại bước (2.3.b), sử dụng $t_{xy}^{(t)}$ đã được xác định tại bước (2.3.a) ta tìm được $K_x^{(t)}$ là tập láng giềng của sản phẩm t_x tại bước lặp thứ t . Tại bước (2.3.c), sử dụng $K_x^{(t)}$ đã xác định tại bước (2.3.b) ta dự đoán được $r_{ix}^{(t)}$ là quan điểm chắc chắn của người dùng u_i cho các sản phẩm t_x tại bước lặp thứ t. Sau bước 2.3, thuật toán kiểm tra nếu thỏa mãn điều kiện hội tụ là không có nhãn phân loại nào được bổ sung vào ma trận dự

đoán, khi đó $r_{ix}^{(t)} = r_{ix}^{(t-1)}$ thì dừng lại, nếu chưa thì quá trình đồng huấn luyện tiếp theo được thực hiện ở vòng lặp tiếp theo.

C. Sinh tư vấn

Sau khi kết thúc quá trình đồng huấn luyện, hệ thống thu được ma trận dự đoán $R_o{}^{(t)}$ là cơ sở để sinh ra tư vấn sản phẩm phù hợp cho người dùng theo ngữ cảnh, điều này được miêu tả cụ thể trong Thuật toán 2 dưới đây.

Như vậy, trên cơ sở bộ khung triển khai phương pháp lọc trước theo ngữ cảnh dựa vào đồng huấn luyện với 3 bước thực hiện ở Mục A, Mục B, Mục C trình bày ở trên, tác giả đề xuất hai thuật toán mới cho lọc trước ngữ cảnh dựa vào đồng huấn luyện theo người dùng (IS-CoTraining-UserItem) và lọc trước ngữ cảnh dựa vào đồng huấn luyện theo sản phẩm (IS-CoTraining-ItemUser) dưới đây.

Thuật toán 2. *Thuật toán IS-CoTraining-UserItem*

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 thủ tục phân tách sản phẩm theo ngữ cảnh (Mục A).

Bước 2. Học dữ liệu bằng mô hình đồng huấn luyện (Theo Thuật toán 1)

Bước 3. *Sinh tư vấn cho người dùng hiện thời u_a trong ngữ cảnh c.*

- Từ ma trận $R_o{}^{(t)}$ sắp xếp các sản phẩm chưa được đánh giá ban đầu bởi người dùng hiện thời u_a theo thứ tự giảm dần của $r_{ix}^{(t)}$. Với mỗi người dùng hiện thời u_a, chọn K_1 sản phẩm có đầu tiên trong số đó tư vấn cho người dùng 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.

Phương pháp lọc trước ngữ cảnh dựa vào sản phẩm đề xuất (IS-CoTraining-ItemUser) có cơ chế thực hiện tương tự phương pháp lọc trước ngữ cảnh dựa vào người dùng (IS-CoTraining-UserItem). Điểm khác biệt cơ bản giữa hai phương pháp đồng huấn luyện 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.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu thực nghiệm

Để thấy rõ hiệu quả của phương pháp đề xuất, tác giả thực hiện tiến hành thực nghiệm trên hai bộ dữ liệu DepaulMovie, InCarMusic [24].

- 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 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.

B. Cài đặt thực nghiệm

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 $Precision@N$, $MAP@N$. Trong bài báo này, tác giả tập trung đánh giá hiệu quả tư vấn danh sách sản phẩm của phương pháp đề xuất trong sự so sánh với các phương pháp tư vấn theo ngữ cảnh cơ sở nên độ đo $Precision@N, MAP@N$ sẽ được lựa chọn để đánh giá kết quả. Cụ thể độ đo như sau:

- Độ chính xác Precision@N cho biết tỷ lệ dự đoán chính xác trong top-N sản phẩm dự đoán cho mỗi người dùng (top-N items).

$$\text{Precision@N} = \frac{|\{relevant\ items\} \cap \{top{-}N\ items\}|}{N} \quad (5)$$

- MAP thể hiện tính đúng đắn về thứ hạng của những gợi ý. $MAP@k$ được định nghĩa thông qua $AP@k$ ($Average\ Precision$) dưới đây.

$$AP@k = \frac{1}{m}\sum_{i=1}^{k}(Precision@i\ nếu\ sản\ phẩm\ thứ\ i\ phù\ hợp) = \frac{1}{m}\sum_{i=1}^{k} Precision@i.rel(i) \quad (6)$$

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}.

$$MAP@k = \frac{1}{|U_{test}|}\sum_{i=1}^{|U_{test}|}(AP@k)_{u_i} \quad (7)$$

2) *Phương pháp thực nghiệm*

Để đánh giá độ chính xác của danh sách sản phẩm tư vấn, tác giả thực hiện phân chia tập dữ liệu U thành 2 tập U_{train} và U_{test} sử dụng 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, tác giả 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.

3) *Các phương pháp tư vấn được sử dụng để so sánh*

- *UserSplitting-BiasedMF* [11]: 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) [25].
- *ItemSplitting-BiasedMF* [8][26]: 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* [26]: 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.
- *ItemSplitting-CoTraining-UserItem*: Phương pháp lọc trước ngữ cảnh dựa vào đồng huấn luyện theo người dùng,kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh nguyên thủy; 2) Lọc cộng tác bằng đồng huấn luyện theo người dùng.
- *ItemSplitting-CoTraining-ItemUser*: Phương pháp lọc trước ngữ cảnh dựa vào đồng huấn luyện theo sản phẩm,kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh nguyên thủy; 2) Lọc cộng tác bằng đồng huấn luyện theo sản phẩm.
- *IS-CoTraining-UserItem*: Phương pháp lọc trước ngữ cảnh dựa vào đồng huấn luyện theo người dùng đề xuất, kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh cải tiến; 2) Lọc cộng tác bằng đồng huấn luyện theo người dùng.
- *IS-CoTraining-ItemUser*: Phương pháp lọc trước ngữ cảnh dựa vào đồng huấn luyện theo sản phẩm đề xuất, kết hợp hai phương pháp: 1) Lọc trước theo ngữ cảnh cải tiến; 2) Lọc cộng tác bằng đồng huấn luyện theo sản phẩm.

C. *Kết quả thực nghiệm*

Kết quả thực nghiệm được đưa ra trong Bảng 1, Bảng 2 nhằm đánh giá một số kịch bản sau:

- So sánh *ItemSplitting-CoTraining-UserItem*, *ItemSplitting-CoTraining-ItemUser* với các phương pháp lọc trước sử dụng 1 chiều ngữ cảnh.
- So sánh *IS-CoTraining-UserItem*, *IS-CoTraining-ItemUser với ItemSplitting-CoTraining-UserItem*, *ItemSplitting-CoTraining-ItemUser*.

Bảng 3. *Giá trị Precision@10, MAP@10 trên tập DepaulMovie*

Phương pháp	Precision@10	MAP@10
UserSplitting-BiasedMF	0.089	**0.161**
ItemSplitting-BiasedMF	0.086	0.147
UISplitting-BiasedMF	0.084	0.144
ItemSplitting-CoTraining-UserItem	0.119	0.135
ItemSplitting - CoTraining-ItemUser	0.121	0.152
IS-CoTraining-UserItem	0.119	0.160
IS-CoTraining-ItemUser	**0.122**	0.159

Bảng 4. *Giá trị Precision@10, MAP@10 trên tập InCarMusic*

Phương pháp	Precision@10	MAP@10
UserSplitting-BiasedMF	0.033	0.125
ItemSplitting-BiasedMF	0.034	0.127
UISplitting-BiasedMF	0.033	0.117
ItemSplitting-CoTraining-UserItem	0.036	0.065
ItemSplitting-CoTraining-ItemUser	0.037	0.112

IS-CoTraining-UserItem	0.037	**0.145**
IS-CoTraining-ItemUser	**0.038**	0.141

Một số nhận xét được đưa ra căn cứ vào phân tích kết quả thực nghiệm đưa ra trong Bảng 1, Bảng 2 như sau:

1) Các phương pháp lọc trước sử dụng 1 chiều ngữ cảnh *ItemSplitting-CoTraining-UserItem*, *ItemSplitting-CoTraining-ItemUser* cho lại Precision@10 tốt hơn, nhưng MAP@10 lại cho kết quả thấp hơn các phương pháp tư vấn theo ngữ cảnh cơ sở cùng hướng. 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 đồng huấn luyện cho lọc cộng tác chưa hẳn là giải pháp tối ưu.

2) Kết hợp phương pháp phân tách sản phẩm theo ngữ cảnh cải tiến (sử dụng đồng thời nhiều chiều ngữ cảnh) và phương pháp *CoTraining-UserItem*, *CoTraining-ItemUser* để tạo thành phương pháp đề xuất ***IS-CoTraining-UserItem***, ***IS-CoTraining-ItemUser***. So sánh về giá trị Precision@10 nhận thấy phương pháp *CoTraining-UserItem, CoTraining-ItemUser* cho Precision@10 cao hơn chút ít so với *ItemSplitting-CoTraining-UserItem, ItemSplitting-CoTraining-ItemUser*. So sánh về giá trị MAP@10 của phương pháp đề xuất lớn hơn *ItemSplitting-CoTraining-UserItem, ItemSplitting-CoTraining-ItemUser* trong cả 2 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-CoTraining-UserItem, IS-CoTraining-ItemUser* cho lại độ chính xác $Precision@10$ tốt hơn các phương pháp cơ sở. Đặc biệt, phương pháp *IS-CoTraining-ItemUser* cho $Precision@10$ cao nhất đối với cả hai tập dữ liệu. Phương pháp *IS-CoTraining-UserItem* 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 hơn trong hai tập dữ liệu. Các kết quả này đưa ra bằng chứng cho thấy phương pháp đề xuất bởi bài báo 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.

3) Trong hai phương pháp đề xuất bởi bài báo, *IS-CoTraining-ItemUser* cho độ chính xác $Precision@10$ cao hơn *IS-CoTraining-UserItem*, 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.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã trình bày đề xuất một phương pháp lọc trước theo ngữ cảnh mới cho hệ tư vấn, cho phép tích hợp đầy đủ thông tin ngữ cảnh và giải quyết hiệu quả vấn đề dữ liệu thưa. Trong đó, việc tích hợp ngữ cảnh được thực hiện bằng thủ tục phân tách sản phẩm theo ngữ cảnh cải tiến. Quá trình phân tách sản phẩm theo ngữ cảnh sẽ biến đổi ma trận đánh giá đa chiều vốn dĩ đã thưa về ma trận đánh giá hai chiều càng trở lên thưa thớt hơn nữa. Để giải quyết vấn đề thưa thớt dữ liệu này, tác giả tiếp cận phương pháp đồng huấn luyện cho lọc cộng tác, 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 một số bộ dữ liệu thực cho thấy phương pháp đề xuất cải thiện đáng kể chất lượng dự đoán so với các phương pháp tư vấn dựa vào ngữ cảnh cơ sở trước đây.

Trong thời gian tới, tác giả dự định sẽ mở rộng nghiên cứu của mình cho hệ tư vấn lai theo ngữ cảnh nhằm tích hợp được nhiều thông tin phục vụ cho quá trình huấn luyện nâng cao chất lượng tư vấn. Ngoài ra tác giả cũng có kế hoạch nghiên cứu phát triển các phương pháp mô hình hóa ngữ cảnh phụ thuộc áp dụng cho hệ tư vấn theo ngữ cảnh.

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, doi: 10.1109/TKDE.2005.99.

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

[3] L. Baltrunas, B. Ludwig, and F. Ricci, "Matrix Factorization Techniques for Context Aware," *Acm Rs*, no. October, pp. 301–304, 2011, doi: 10.1145/2043932.2043988.

[4] 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, doi: 10.1080/13658816.2017.1400550.

[5] 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, doi: 10.1016/j.compeleceng.2018.01.034.

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

[7] 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, doi: 10.1007/s11257-012-9135-y.

[8] 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.

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

[10] M. Unger, A. Bar, B. Shapira, and L. Rokach, "Towards

latent context-aware recommendation systems," *Knowledge-Based Syst.*, vol. 104, pp. 165–178, 2016, doi: 10.1016/j.knosys.2016.04.020.
[11] 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, doi: 10.1145/2554850.2554989.
[12] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse Recommendation: N-dimensional Tensor Factorization for Context-aware Collaborative Filtering," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010, pp. 79–86, doi: 10.1145/1864708.1864727.
[13] 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, doi: 10.1145/1055709.1055714.
[14] Y. Zheng, "Tutorial : Context In Recommender Systems," 2016.
[15] 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, [Online]. Available: https://dl.acm.org/citation.cfm?id=295795.
[16] 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, doi: 10.1145/1363686.1363903.
[17] 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, [Online]. Available: https://dl.acm.org/citation.cfm?id=657311.
[18] N. D. Phuong and T. M. Phuong, "Collaborative Filtering by Multi-task Learning," vol. 00, no. c, pp. 1–6, 2008.
[19] 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, doi: 10.1016/j.eswa.2017.12.020.
[20] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 2014.
[21] A. Z. Olivier Chapelle, Bernhard Scho¨lkopf, *A semi-supervised learning*, vol. 1, no. 2. The MIT Press Cambridge, Massachusetts London, England, 2009.
[22] P. Rai, "Semi-supervised Learning," in *CS 5350/6350: Machine Learning*, 2011, vol. 2011.
[23] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Adv. Artif. Intell.*, vol. 2009, 2009, doi: 10.1155/2009/421425.
[24] 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, doi: 10.1109/ICDMW.2015.222.
[25] 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, doi: 10.1109/MC.2009.263.
[26] 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, doi: 10.1007/s11257-012-9137-9.

A CONTEXTUAL PRE-FILTERING METHOD FOR CONTEXT-AWARE RECOMMENDER SYSTEMS

Abstract: Recommender systems are specially designed to toward users to interested items when huge information from Internet is beyond the user's processing capability. A common characteristic of recommender systems is that they mainly focus on modeling users, items and ratings. In parallel, there is an understanding that it is also important to consider the context in which a recommendation is made. Although the integration of context into recommender systems has been shown to improve quality of suggestions significantly, the main difficulty of context-aware recommender systems is how to integrate effectively and the data sparseness problem which directly affect to quality of the recommendation. In this paper, I will propose a new contextual pre-filtering method that allow fully integrated context situations and resolve effectively the data sparseness problem. In there, the contextual integration is done by a item splitting proceduce based on context and the data sparseness issue is resolved through the training process according to the co-training model for classification problem of collaborative filtering. The experimental results on some real data sets show that the proposed method outperforms several baselines and state-of-the-art context-aware recommendation methods.

Keyword: Context-aware recommender system - CARS; Context-aware collaborative filtering - CACF; Context; Contextual pre-filtering; Item splitting; Co-training model.

Đỗ Thị Liên, Nhận bằng tốt nghiệp đại học, thạc sỹ và học vị tiến sĩ tại Học viện Công nghệ Bưu chính Viễn thông vào các năm 2010, 2013, 2020. Hiện là giảng viên tại Học Viện Công nghệ Bưu Chính Viễn Thông.

Lĩnh vực nghiên cứu chính: học máy ứng dụng trong lọc thông tin, phát triển ứng dụng đa phương tiện.