Hợp nhất lọc cộng tác và lọc nội dung bằng phương pháp học bán giám sát

Đỗ Thị Liên, Nguyễn Duy Phương, Từ Minh Phương
Học viện Công nghệ Bưu chính Viễn thông
E-mail: liendt@ptit.edu.vn, phuongnd@ptit.edu.vn, phuongtm@ptit.edu.vn
Tác giả liên hệ: Đỗ Thị Liên
Ngày nhận: 26/02/2017, ngày sửa chữa: 06/03/2017, ngày duyệt đăng: 10/07/2017

Tóm tắt: Hệ tư vấn là hệ thống tự động cung cấp thông tin phù hợp và gỡ bỏ thông tin không phù hợp cho mỗi người dùng. Hệ tư vấn được xây dựng dựa trên hai kỹ thuật lọc thông tin chính: lọc cộng tác và lọc nội dung. Lọc nội dung thực hiện hiệu quả trên các loại tài liệu văn bản nhưng gặp phải vấn đề trích chọn đặc trưng trên các dạng thông tin đa phương tiện. Lọc cộng tác thực hiện tốt trên tất cả các dạng thông tin nhưng gặp phải vấn đề dữ liệu thưa, người dùng mới và sản phẩm mới. Trong bài báo này, chúng tôi đề xuất một phương pháp hợp nhất giữa lọc cộng tác và lọc nội dung bằng phương pháp đồng huấn luyện. Kết quả thử nghiệm trên các bộ dữ liệu thực tế cho thấy phương pháp đề xuất tận dụng hiệu quả ưu điểm và hạn chế đáng kể nhược điểm của mỗi phương pháp lọc truyền thống.

Từ khóa: *Lọc cộng tác, lọc nội dung, lọc kết hợp, đồng huấn luyện, học có giám sát, học không giám sát, học bán giám sát.*

Title: **Unifying Collaborative and Content-based Filtering by Semi-Supervised Learning**

Abstract: A recommender system is an automated system that provides appropriate information and removing inappropriate information for users. It is based on two main information filtering techniques: collaborative filtering and content-based filtering. Content-based filtering performs well with information in text form but has difficulty in feature selection with multimedia information. Collaborative filtering performs well on all types of information but has problems with sparse data, new users, and new items. In this paper, we propose a new model that unifies collaborative filtering and content-based filtering by a co-training method. Experimental results on real datasets showed that the proposed method effectively makes use of the advantages of state-of-the-art filtering methods and significantly overcomes their disadvantages.

Keywords: *Collaborative filtering, content-based filtering, hybrid filtering, co-training, supervised learning, unsupervised learning, semi-supervised learning.*

I. GIỚI THIỆU

Người dùng sử dụng các dịch vụ Internet trực tuyến hiện nay luôn trong tình trạng quá tải thông tin. Để tiếp cận được thông tin hữu ích, người dùng thường phải xử lý, loại bỏ phần lớn thông tin không cần thiết. Hệ tư vấn (recommender systems) cung cấp một giải pháp nhằm giảm tải thông tin bằng cách dự đoán và cung cấp một danh sách ngắn các sản phẩm (trang web, bản tin, phim, video, v.v.) phù hợp cho mỗi người dùng. Hệ tư vấn được xây dựng dựa trên một tập gồm N người dùng, $U = \{u_1, u_2, \ldots, u_N\}$, và $P = \{p_1, p_2, \ldots, p_M\}$, là một tập gồm M sản phẩm. Mỗi sản phẩm $p_x \in P$ 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. Để thuận tiện trong trình bày, ta viết $p_x \in P$ ngắn gọn thành $x \in P$; và $u_i \in U$ là $i \in U$. 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 $i = 1, 2, \ldots, N$; $x = 1, 2, \ldots, M$. Giá trị r_{ix} thể hiện đánh giá của người dùng $i \in U$ cho một sản phẩm $x \in P$. Thông thường r_{ix} nhận một giá trị thuộc một miền $F = \{1, 2, \ldots, g\}$, được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế ghi nhận phản hồi của người dùng. Giá trị $r_{ix} = 0$ được hiểu là người dùng i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm x. Ma trận đánh giá của các hệ thống tư vấn thực tế thường rất thưa. Mật độ các giá trị $r_{ix} \neq 0$ thường nhỏ hơn 1%, nghĩa là hầu hết các giá trị r_{ix} là 0 [1, 2]. Ma trận R chính là đầu vào của các hệ thống tư vấn cộng tác [3].

Mỗi sản phẩm $x \in P$ được biểu diễn thông qua $|C|$ đặc trưng nội dung, biểu diễn bởi tập $C = \{c_1, c_2, \ldots, c_{|C|}\}$. Các đặc trưng $s \in C$ có được từ các phương pháp trích chọn đặc trưng (feature extraction) trong lĩnh vực truy vấn

thông tin. Ví dụ, $x \in P$ là một phim thì các đặc trưng nội dung biểu diễn một phim có thể là $C = \{$*thể loại phim*, *nước sản xuất*, *hãng phim*, *diễn viên*, *đạo diễn*$\ldots\}$. Gọi $w_i = [w_{i1}, w_{i2}, \ldots, w_{i|C|}]$ là véc tơ trọng số các giá trị đặc trưng nội dung của sản phẩm s đối với mỗi người dùng $i \in U$. Khi đó, ma trận trọng số $W = [w_{is}]$, với $i = 1, 2, \ldots, N$, $s = 1, 2, \ldots, |C|$, là đầu vào của các hệ thống tư vấn theo nội dung sản phẩm [2, 4].

Mỗi người dùng $i \in U$ được biểu diễn thông qua tập $T = \{t_1, t_2, \ldots, t_{|T|}\}$, bao gồm $|T|$ đặc trưng nội dung. Các đặc trưng $q \in T$ thông thường là thông tin cá nhân của mỗi người dùng (demographic information). Ví dụ, $i \in U$ là một người dùng thì các đặc trưng nội dung biểu diễn người dùng i có thể là $T = \{$*giới tính*, *độ tuổi*, *nghề nghiệp*, *trình độ*,$\ldots\}$. Gọi $v_x = [v_{x1}, v_{x2}, \ldots, v_{x|T|}]$ là véc tơ trọng số biểu diễn các giá trị đặc trưng nội dung người dùng $q \in T$ đối với mỗi sản phẩm $x \in P$. Khi đó, ma trận trọng số $V = [v_{xq}]$, với $x = 1, 2, \ldots, M$; $q = 1, 2, \ldots, |T|$, là đầu vào của các hệ thống tư vấn theo nội dung thông tin người dùng [2, 5].

Tiếp đến ta ký hiệu, $P_i \subseteq P$ là tập các sản phẩm $x \in P$ được đánh giá bởi người dùng $i \in U$ và $U_x \subseteq U$ là tập các người dùng đã đánh giá sản phẩm $x \in P$. Với một người dùng cần được tư vấn $j \in U$ (được gọi là người dùng hiện thời, người dùng cần được tư vấn, hay người dùng tích cực), nhiệm vụ của các phương pháp tư vấn là gợi ý K sản phẩm $x \in (P \setminus P_j)$ phù hợp nhất đối với người dùng j [3, 6].

Đã có nhiều đề xuất khác nhau giải quyết bài toán tư vấn. Tuy vậy, ta có thể phân loại thành ba hướng tiếp cận chính: tư vấn theo nội dung, tư vấn cộng tác và tư vấn kết hợp [1, 3, 7]. Hệ tư vấn theo nội dung xây dựng phương pháp dự đoán dựa trên ma trận trọng số các đặc trưng nội dung sản phẩm $W = [w_{is}]$ hoặc ma trận trọng số các đặc trưng nội dung người dùng $V = [v_{xq}]$ [2, 4, 8]. Lọc nội dung thực hiện khá tốt trên các loại thông tin văn bản nhưng gặp khó khăn trong trích chọn đặc trưng các sản phẩm đa phương tiện (ví dụ hình ảnh, âm thanh, v.v.). Một người dùng mới tham gia hệ thống sẽ có hồ sơ sử dụng sản phẩm là tập rỗng ($\emptyset$). Khi đó, hệ thống sẽ không thể gợi ý được các sản phẩm phù hợp với người dùng này [1, 8].

Hệ tư vấn cộng tác xây dựng phương pháp dự đoán dựa trên ma trận đánh giá $R = [r_{ix}]$ [3, 8–10]. Trong đó, giá trị r_{ix} phản ánh quan điểm của người dùng $i \in U$ đối với các sản phẩm $x \in P$. Lọc cộng tác thực hiện tốt trên tất cả các loại thông tin, đặc biệt đối với thông tin đa phương tiện (ví dụ hình ảnh, âm thanh, v.v.). Chính vì lý do này, lọc cộng tác được sử dụng rộng rãi hơn lọc nội dung trong các hệ thống thương mại điện tử [8]. Thách thức lớn nhất của lọc cộng tác là vấn đề dữ liệu thưa, người dùng mới và sản phẩm mới [1, 3].

Hệ tư vấn kết hợp xây dựng phương pháp dự đoán dựa trên cả ba ma trận R, W, V [2, 6, 11]. Hệ tư vấn kết hợp được tiếp cận theo bốn xu hướng chính: kết hợp tuyến tính giữa lọc cộng tác và lọc nội dung, kết hợp các đặc trưng của lọc cộng tác vào lọc nội dung, kết hợp các đặc trưng của lọc nội dung vào lọc cộng tác và xây dựng mô hình hợp nhất cho cả hai phương pháp lọc [2]. Hai vấn đề cơ bản cần giải quyết đối với phương pháp tiếp cận này là tìm ra phép biểu diễn hợp lý giữa đánh giá người dùng của lọc cộng tác với các đặc trưng của lọc nội dung và phương pháp dự đoán chung cho cả hai phương pháp [1, 8].

Trong bài báo này, chúng tôi đề xuất một mô hình hợp nhất giữa lọc cộng tác và lọc nội dung bằng phương pháp học bán giám sát nhằm tận dụng lợi thế và hạn chế khó khăn của mỗi phương pháp lọc. Phương pháp được xây dựng dựa trên cơ sở xây dựng mô hình hợp nhất giữa đánh giá người dùng của lọc cộng tác và hồ sơ người dùng của lọc nội dung để thống nhất các mô hình dự đoán dựa vào người dùng. Tiếp đến, chúng tôi xây dựng mô hình hợp nhất giữa đánh giá sản phẩm của lọc cộng tác và hồ sơ sản phẩm của lọc nội dung để thống nhất các mô hình dự đoán dựa vào sản phẩm. Cuối cùng, chúng tôi xây dựng mô hình học bán giám sát để hợp nhất cả hai phương pháp dự đoán dựa vào người dùng và phương pháp dự đoán dựa vào sản phẩm.

Bài báo có cấu trúc như sau: Mục II trình bày phương pháp ước lượng trọng số các đặc trưng nội dung người dùng và sản phẩm của lọc nội dung; Mục III trình bày phương pháp học bán giám sát dựa vào đánh giá người dùng, đặc trưng sản phẩm và đặc trưng người dùng; Mục IV trình bày phương pháp thử nghiệm và đánh giá; Mục V là kết luận và hướng phát triển tiếp theo của bài báo.

II. HỢP NHẤT BIỂU DIỄN GIÁ TRỊ CÁC ĐẶC TRƯNG NỘI DUNG

Như đã giới thiệu ở trên, bài toán tư vấn kết hợp thực hiện dự đoán dựa trên tập đánh giá của người dùng đối với sản phẩm, cùng với tập đặc trưng nội dung sản phẩm và đặc trưng người dùng. Trong mục này, chúng tôi trình bày đề xuất phương pháp hợp nhất biểu diễn giá trị các đặc trưng nội dung vào ma trận đánh giá của lọc cộng tác. Đây cũng là bước đầu tiên trong xây dựng mô hình học bán giám sát cho hệ tư vấn kết hợp.

Không hạn chế tính tổng quát của bài toán phát biểu trong mục I, ta giả thiết giá trị đánh giá của người dùng $i \in U$ đối với sản phẩm $x \in P$ được xác định theo công thức:

$$r_{ix} = \begin{cases} v, & \text{nếu người dùng } i \text{ đánh giá sản phẩm } x \text{ là } v, \\ 0, & \text{nếu người dùng } i \text{ chưa đánh giá sản phẩm } x. \end{cases} \quad (1)$$

Bảng I
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 II
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 III
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

Mỗi sản phẩm $x \in P$ được biểu diễn thông qua tập $C = \{c_1, c_2, \ldots, c_{|C|}\}$, bao gồm $|C|$ đặc trưng nội dung, được xác định theo công thức:

$$c_{xs} = \begin{cases} 1, & \text{nếu sản phẩm } x \text{ có đặc trưng } s, \\ 0, & \text{nếu sản phẩm } x \text{ không có đặc trưng } s. \end{cases} \quad (2)$$

Mỗi người dùng $i \in U$ được biểu diễn thông qua tập $T = \{t_1, t_2, \ldots, t_{|T|}\}$, bao gồm $|T|$ đặc trưng nội dung, được xác định theo công thức:

$$t_{iq} = \begin{cases} 1, & \text{nếu người dùng } i \text{ có đặc trưng } q, \\ 0, & \text{nếu người dùng } i \text{ không có đặc trưng } q. \end{cases} \quad (3)$$

Ví dụ, với hệ gồm 3 người dùng, $U = \{u_1, u_2, u_3\}$, và 4 sản phẩm, $P = \{p_1, p_2, p_3, p_4\}$. Ma trận đánh giá R được cho trong Bảng I; Ma trận đặc trưng nội dung sản phẩm C được cho trong Bảng II; Ma trận đặc trưng nội dung người dùng T được cho trong Bảng III. Hệ tư vấn cộng tác được xây dựng dựa trên ma trận đánh giá R [9, 12]. Hệ tư vấn nội dung được xây dựng dựa trên ma trận các đặc trưng nội dung C và T [4, 5]. Hệ tư vấn lai xây dựng dựa trên cả ba ma trận R, C và T [2, 13].

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: xác định tập sản phẩm người dùng đã từng truy cập hay sử dụng trong quá khứ và ước lượng trọng số mỗi đặc trưng nội dung sản phẩm trong hồ sơ người dùng [2, 4, 8]. Gọi $P_i \subseteq P$, được xác định theo công thức:

$$P_i = \{ x \in P | \, r_{ix} \neq 0 \ (i \in U) \}, \quad (4)$$

là tập sản phẩm người dùng $i \in U$ đã đánh giá. 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 $s \in C$ đối với mỗi hồ sơ người dùng $i \in U$.

Gọi $Item(i, s)$ là tập các sản phẩm trong P_i chứa đựng đặc trưng $s \in C$ được xác định theo công thức:

$$Item(i, s) = \{ x \in P_i | \, c_{xs} \neq 0 \ (i \in U, s \in C) \}. \quad (5)$$

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

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 s đối với người dùng i. Phương pháp phổ dụng nhất được sử dụng trong xây dựng hồ sơ người dùng là kỹ thuật tf-idf [4, 8]. 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 chúng tôi 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 [7, 9, 10], 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, chúng tôi 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, chúng tôi 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 $s \in C$ đối với người dùng $i \in U$ 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 thử nghiệm, chúng tôi tính toán số lượng trung bình của tất cả người dùng đã đánh giá các sản phẩm $x \in P$. Sau đó, chọn θ tương đương với $2/3$ số lượng trung bình các đánh giá của tập người dùng đã đánh giá sản phẩm $x \in P$ chứa đựng đặc trưng $s \in C$. 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.

Bảng IV
MA TRẬN HỒ SƠ NGƯỜI DÙNG w_{is}

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

Bảng V
MA TRẬN ĐÁNH GIÁ MỞ RỘNG 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

Giá trị w_{is}, được ước lượng theo công thức:

$$w_{is} = \begin{cases} \dfrac{1}{|Item(i,s)|} \displaystyle\sum_{x \in Item(i,s)} r_{ix}, & \text{nếu } |Item(i,s)| \geq \theta, \\ \dfrac{1}{\theta} \displaystyle\sum_{x \in Item(i,s)} r_{ix}, & \text{nếu } |Item(i,s)| < \theta, \end{cases} \tag{6}$$

phản ánh quan điểm của người dùng $i \in U$ đối với các đặc trưng nội dung sản phẩm $s \in C$ trong quá khứ. Dễ dàng nhận thấy $w_{is} \in F$, trong đó $F = \{1, 2, \ldots, g\}$. 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, chúng tôi 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 công thức:

$$r_{ix} = \begin{cases} r_{ix}, & \text{nếu } x \in P, \\ w_{is}, & \text{nếu } s \in C \ (x = s), \end{cases} \tag{7}$$

trong đó $x = s$ ($s \in C$) đóng vai trò như một sản phẩm phụ bổ sung vào ma trận đánh giá về phía sản phẩm.

Ví dụ với hệ có ma trận đánh giá theo Bảng I, ma trận đặc trưng sản phẩm theo Bảng II, ma trận đặc trưng người dùng theo Bảng III, chọn $\theta = 2$, khi đó ta sẽ tính toán được tập hồ sơ người dùng $\{w_{is} | i \in U, s \in C\}$ trong Bảng IV và ma trận đánh giá mở rộng theo (7) trong Bảng V.

Hệ tư vấn được xác định theo (7) 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 [2, 6, 8]. 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 dựa vào (7) đều cho lại kết quả không cao. Vấn đề này sẽ được chúng tôi giải quyết trong mục tiếp theo của bài báo.

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 [2]. Gọi $U_x \subseteq U$, được xác định theo công thức:

$$U_x = \{ i \in U | \, r_{ix} \neq 0 \ (x \in P) \}, \tag{8}$$

là tập người dùng thuộc U đã sử dụng sản phẩm $x \in P$. 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 $q \in T$ đối với mỗi hồ sơ sản phẩm $x \in P$.

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

$$User(x, q) = \left\{ i \in U_x | \, t_{iq} \neq 0 \ (x \in P, q \in T) \right\}. \tag{9}$$

Khi đó, $|User(x, q)|$ chính là số lần sản phẩm $x \in P$ được tập người dùng có đặc trưng nội dung $q \in T$ sử dụng trong quá khứ.

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, chúng tôi 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, chúng tôi 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 $q \in T$ đối với sản phẩm $x \in P$ là ν_{xq} 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 ν_{xq} được tính bằng tổng của tất cả các giá trị đánh giá chia cho θ.

Giá trị ν_{qx}, được ước lượng theo công thức:

$$\nu_{qx} = \begin{cases} \dfrac{1}{|User(x,q)|} \displaystyle\sum_{i \in User(x,q)} r_{ix}, & \text{nếu } |User(x,q)| \geq \theta, \\ \dfrac{1}{\theta} \displaystyle\sum_{i \in User(x,q)} r_{ix}, & \text{nếu } |User(x,q)| < \theta, \end{cases} \tag{10}$$

biểu diễn hồ sơ sản phẩm $x \in P$ đã được tập những người dùng chứa đựng đặc trưng $q \in T$ 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

Bảng VI
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 VII
MA TRẬN ĐÁNH GIÁ MỞ RỘNG r_{ix} THEO HỒ SƠ SẢN PHẨM

	p_1	p_2	p_3	p_4
u_1	6	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

nhận xét này, chúng tôi 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:

$$r_{ix} = \begin{cases} r_{ix}, & \text{nếu } i \in U \text{ và } r_{ix} \neq 0, \\ v_{qx}, & \text{nếu } q \in T \text{ và } v_{qx} \neq 0 \ (i = q), \end{cases} \quad (11)$$

trong đó, $i = q$ ($q \in T$) đóng vai trò như một người dùng phụ bổ sung vào để mở rộng ma trận đánh giá về phía người dùng.

Ví dụ với hệ có ma trận đánh giá theo Bảng I, ma trận đặc trưng người dùng theo Bảng III, chọn $\theta = 2$, khi đó ta sẽ tính toán được tập hồ sơ sản phẩm $\{v_{qx} | x \in P, q \in T\}$ trong Bảng VI và ma trận đánh giá mở rộng về phía người dùng theo (11) trong Bảng VII.

Hệ tư vấn được xác định theo (11) 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 [2, 10]. 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 (11) đều cho lại kết quả không cao. Vấn đề này sẽ được chúng tôi giải quyết trong mục tiếp theo của bài báo.

III. MÔ HÌNH HỌC BÁN GIÁM SÁT CHO LỌC KẾT HỢP

Như đã đề cập ở trên, các phương pháp tư vấn dựa vào các công thức (7) và (11) đều gặp phải vấn đề dữ liệu thưa [2, 3]. Để khắc phục điều này, chúng tôi đề xuất thuật toán tư vấn kết hợp bằng phương pháp học bán giám sát. Thuật toán được xây dựng dựa trên hai thủ tục bán giám sát: bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm và bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng. Bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm cho phép ta phát hiện ra những sản phẩm mới có khả năng cao phù hợp cho mỗi người dùng. Những sản phẩm mới được phát hiện sẽ chuyển giao cho quá trình bán giám sát theo đánh giá sản phẩm cùng tập đặc trưng người dùng. Ngược lại, thủ tục bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng cho phép ta phát hiện ra những người dùng mới có khả năng phù hợp cao đối với sản phẩm. Những người dùng mới được dự đoán sẽ được chuyển giao cho quá trình bán giám sát theo tập đánh giá người dùng cùng tập đặc trưng sản phẩm. Hai quá trình bán giám sát được thực hiện đồng thời và bổ sung qua lại các giá trị dự đoán chắc chắn cho nhau để nâng cao chất lượng tư vấn.

1. Bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm

Để hạn chế ảnh hưởng của vấn đề dữ liệu thưa, với mỗi người dùng $i \in U$ chúng tôi xây dựng tập S_i, được định nghĩa theo công thức:

$$S_i = \{j \in U | \left|P_i \cap P_j\right| \geq \theta_1 \text{ và } \left|C_i \cap C_j\right| \geq \theta_2\}, \quad (12)$$

để giám sát việc tính toán mức độ tương tự giữa các cặp người dùng. Trong công thức (12), P_i được xác định theo (4), C_i được xác định bởi

$$C_i = \{s \in C | r_{is} \neq 0\}. \quad (13)$$

S_i được xác định theo (12) là tập người dùng thuộc U có số lượng đánh giá giao nhau với người dùng 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 [7, 8], chúng tôi bán giám sát việc tính toán mức độ tương tự giữa các cặp người dùng của lọc cộng tác theo công thức (14), bán giám sát việc tính toán mức độ tương tự giữa các cặp người dùng của lọc nội dung theo công thức (15), bán giám sát việc tính toán mức độ tương tự giữa các cặp người dùng của lọc kết hợp theo công thức (16) (xem đầu trang sau).

Trong các công thức (14), (15), (16), P_i được xác định theo công thức (4), C_i được xác định theo công thức (13);

$$a_{ij} = \begin{cases} \dfrac{\sum\limits_{x \in P_i \cap P_j} (r_{ix} - \overline{r}_i)(r_{jx} - \overline{r}_j)}{\sqrt{\sum\limits_{x \in P_i \cap P_j} (r_{ix} - \overline{r}_i)^2} \sqrt{\sum\limits_{x \in P_i \cap P_j} (r_{jx} - \overline{r}_j)^2}}, & \text{nếu } j \in S_i, \\ 0, & \text{nếu } j \notin S_i, \end{cases} \tag{14}$$

$$b_{ij} = \begin{cases} \dfrac{\sum\limits_{s \in C_i \cap C_j} (r_{is} - \ddot{r}_i)(r_{js} - \ddot{r}_j)}{\sqrt{\sum\limits_{s \in C_i \cap C_j} (r_{is} - \ddot{r}_i)^2} \sqrt{\sum\limits_{s \in C_i \cap C_j} (r_{js} - \ddot{r}_j)^2}}, & \text{nếu } j \in S_i, \\ 0, & \text{nếu } j \notin S_i, \end{cases} \tag{15}$$

$$u_{ij} = \begin{cases} \dfrac{\sum\limits_{x \in H_i \cap H_j} (r_{is} - \overline{\overline{r}}_i)(r_{js} - \overline{\overline{r}}_j)}{\sqrt{\sum\limits_{x \in H_i \cap H_j} (r_{is} - \overline{\overline{r}}_i)^2} \sqrt{\sum\limits_{x \in H_i \cap H_j} (r_{js} - \overline{\overline{r}}_j)^2}}, & \text{nếu } i \in S_i \text{ và } a_{ij} \geq \alpha \text{ và } b_{ij} \geq \alpha \\ 0, & \text{trong các trường hợp khác.} \end{cases} \tag{16}$$

H_i, $\overline{r}_i$, $\ddot{r}_i$, $\overline{\overline{r}}_i$ được xác định tuần tự theo các công thức (17), (18), (19) và (20),

$$H_i = P_i \cup C_i, \tag{17}$$

$$\overline{r}_i = \frac{1}{\left|P_i \cap P_j\right|} \sum_{x \in P_i \cap P_j} r_{ix}, \tag{18}$$

$$\ddot{r}_i = \frac{1}{\left|C_i \cap C_j\right|} \sum_{s \in C_i \cap C_j} r_{is}, \tag{19}$$

$$\overline{\overline{r}}_i = \frac{1}{\left|H_i \cap H_j\right|} \sum_{x \in H_i \cap H_j} r_{ix}. \tag{20}$$

Rõ ràng, a_{ij} được xác định trên S_i theo (14) 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 (15) 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 (16) 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 i, 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 bài báo này, bằng thực nghiệm chúng tôi 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, chúng tôi xây dựng tập láng giềng cho người dùng $i \in U$ theo công thức:

$$K_i = \{j \in S_i | u_{ij} > \alpha\}. \tag{21}$$

Phương pháp dự đoán các sản phẩm mới $x \in P$ chưa được người dùng i biết đến được thực hiện theo công thức: [3, 9]

$$r_{ix} = \overline{\overline{r}}_i + \frac{\sum\limits_{j \in K_i} (r_{jx} - \overline{\overline{r}}_j) u_{ij}}{\sum\limits_{j \in K_i} \left|u_{ij}\right|}. \tag{22}$$

Những sản phẩm mới $x \in P$ có giá trị dự đoán r_{ix} theo (22) là những dự đoán tin cậy được bổ sung vào ma trận đánh giá mở rộng theo hồ sơ sản phẩm để phục vụ quá trình bán giám sát theo tập đánh giá sản phẩm cùng tập đặc trưng người dùng. Phương pháp bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng sẽ được chúng tôi trình bày trong mục tiếp theo của bài báo.

2. Bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng

Tương tự như người dùng, với mỗi sản phẩm $x \in P$, chúng tôi xây dựng tập S_x, được định nghĩa theo công thức:

$$S_x = \{y \in P : \left|U_x \cap U_y\right| \geq \gamma_1 \text{ và } \left|T_x \cap T_y\right| \geq \gamma_2\}, \tag{23}$$

để giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm. Trong công thức (23), U_x được xác định theo công thức (8), T_x được xác định theo công thức:

$$T_x = \{q \in T : r_{qx} \neq 0\}. \tag{24}$$

$$a_{xy} = \begin{cases} \dfrac{\sum\limits_{i \in U_x \cap U_y} (r_{ix} - \overline{r}_x)(r_{iy} - \overline{r}_y)}{\sqrt{\sum\limits_{i \in U_x \cap U_y} (r_{ix} - \overline{r}_x)^2} \sqrt{\sum\limits_{i \in U_x \cap U_y} (r_{iy} - \overline{r}_y)^2}}, & \text{nếu } y \in S_x, \\ 0, & \text{nếu } y \notin S_x, \end{cases} \tag{25}$$

$$b_{xy} = \begin{cases} \dfrac{\sum\limits_{q \in T_x \cap T_y} (r_{qx} - \ddot{r}_x)(r_{qy} - \ddot{r}_y)}{\sqrt{\sum\limits_{q \in T_x \cap T_y} (r_{qx} - \ddot{r}_x)^2} \sqrt{\sum\limits_{q \in T_x \cap T_y} (r_{qy} - \ddot{r}_y)^2}}, & \text{nếu } y \in S_x, \\ 0, & \text{nếu } y \notin S_x, \end{cases} \tag{26}$$

$$p_{xy} = \begin{cases} \dfrac{\sum\limits_{i \in H_x \cap H_y} (r_{ix} - \overline{\overline{r}}_x)(r_{iy} - \overline{\overline{r}}_y)}{\sqrt{\sum\limits_{i \in H_x \cap H_y} (r_{ix} - \overline{\overline{r}}_x)^2} \sqrt{\sum\limits_{i \in H_x \cap H_y} (r_{iy} - \overline{\overline{r}}_y)^2}}, & \text{nếu } y \in S_x \text{ và } a_{xy} \geq \alpha \text{ và } b_{xy} \geq \alpha, \\ 0, & \text{trong các trường hợp khác.} \end{cases} \tag{27}$$

S_x được xác định theo (23) là tập sản phẩm $y \in P$ có số lượng người dùng đánh giá giao nhau với sản phẩm 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, chúng tôi bán giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm của lọc cộng tác theo công thức (25), bán giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm của lọc nội dung theo công thức (26), bán giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm của lọc kết hợp theo công thức (27).

Trong các công thức (25), (26), (27), U_x được xác định theo công thức (8), T_x được xác định theo công thức (24), $H_x, \overline{r}_x, \ddot{r}_x, \overline{\overline{r}}_x$ được xác định tuần tự theo các công thức (28), (29), (30), (31),

$$H_x = U_x \cup T_x, \tag{28}$$

$$\overline{r}_x = \frac{1}{|U_x \cap U_y|} \sum_{i \in U_x \cap U_y} r_{ix}, \tag{29}$$

$$\ddot{r}_x = \frac{1}{|T_x \cap T_y|} \sum_{q \in T_x \cap T_y} r_{qx}, \tag{30}$$

$$\overline{\overline{r}}_x = \frac{1}{|H_x \cap H_y|} \sum_{i \in H_x \cap H_y} r_{ix}. \tag{31}$$

Rõ ràng, a_{xy} được xác định trên S_x theo (25) 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 (26) 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 u_{xy} được xác định theo (27) 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 x, 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 bài báo này, bằng thực nghiệm chúng tôi 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, chúng tôi xây dựng tập láng giềng cho sản phẩm $x \in P$ theo công thức:

$$K_x = \{y \in S_x : p_{xy} > \alpha\}. \tag{32}$$

Phương pháp dự đoán mức độ phù hợp của người dùng $i \in U$ đối với sản phẩm $x \in P$ được thực hiện theo công thức: [3, 7, 10]

$$r_{ix} = \frac{\sum_{y \in K_x} p_{xy} r_{iy}}{\sum_{y \in K_x} |p_{xy}|}. \tag{33}$$

Giá trị dự đoán r_{ix} theo (33) phản ánh mức độ phù hợp của người dùng $i \in U$ đối với sản phẩm $x \in P$ được bổ sung vào ma trận đánh giá mở rộng theo sản phẩm để phục vụ quá trình bán giám sát theo tập đánh giá người dùng và tập đặc trưng sản phẩm. Hai quá trình bán giám sát được thực hiện đồng thời và bổ sung qua lại cho nhau các giá trị dự đoán chắc chắn r_{ix} để nâng cao kết quả tư vấn. Thuật toán học bán giám sát đồng thời trên tập đánh giá người

dùng, đặc trưng sản phẩm, tập đánh giá sản phẩm và đặc trưng người dùng sẽ được chúng tôi trình bày trong mục tiếp theo của bài báo.

3. Thuật toán học bán giám sát cho lọc kết hợp

Như đã được trình bày ở trên, phương pháp bán giám sát theo đánh giá người dùng cùng tập đặc trưng sản phẩm cho phép ta phát hiện những sản phẩm mới phù hợp nhất đối với mỗi người dùng. Phương pháp bán giám sát theo đánh giá sản phẩm cùng tập đặc trưng người dùng cho phép ta phát hiện những người dùng mới phù hợp nhất đối với mỗi sản phẩm. Trong mục này, chúng tôi đề xuất xây dựng thuật toán học bán giám sát đồng thời để xử lý quá trình chuyển giao kết quả dự đoán giữa quá trình bán giám sát từ tập đánh giá người dùng cùng tập đặc trưng sản phẩm đến quá trình bán giám sát từ tập đánh giá sản phẩm cùng tập đặc trưng người dùng, thuật toán đề xuất được mô tả chi tiết trong Thuật toán 1.

Tại bước (2.2), quá trình bán giám sát theo tập đánh giá sản phẩm 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 $q \in U$ đối với sản phẩm $x \in C$ của vòng lặp thứ (t) theo công thức (10). 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 (11). 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 $x \in P$ của vòng lặp thứ (t) theo công thức (23). Sử dụng $s_i^{(t)}$, bước (2.2.d) ta xác định được $P_{xy}^{(t)}$ là mức độ tương tự giữa các cặp sản phẩm $x, y \in P$ trên cả tập đánh giá sản phẩm và tập đặc trưng người dùng của vòng lặp thứ (t) theo công thức (27). 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 x của vòng lặp thứ (t) theo công thức (32). 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 $i \in U$ đối với sản phẩm $x \in P$ của 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 theo tập đánh giá người cùng tập đặc trưng sản phẩm 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 huấn luyện đồng thời 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 x có giá trị $r_{ix}^{(t)}$ lớn nhất tư vấn cho người dùng i.

Thuật toán 1: Thuật toán học bán giám sát

Đầu vào:

Ma trận $R = \{r_{ix}\}$ được xác định theo (1).
Ma trận $C = \{c_{xs}\}$ được xác định theo (2).
Ma trận $T = \{t_{iq}\}$ được xác định theo (3).
Người dùng $i \in U$ là người dùng cần được tư vấn.

Đầu ra:

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

Các bước tiến hành:

begin

Bước 1 *(Khởi tạo):*

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

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

Bước 2 *(Bước lặp):*

repeat

2.1. Bán giám sát tập đánh giá người dùng và tập đặc trưng sản phẩm:

a) Xác định trọng số các đặc trưng nội dung sản phẩm $w_{is}^{(t)}$ tại vòng lặp thứ t theo công thức (6).

b) Mở rộng ma trận đánh giá theo hồ sơ người dùng $r_{ix}^{(t)}$ tại vòng lặp thứ t theo công thức (7).

c) Xác định $S_i^{(t)}$ theo công thức (12).

d) Tính toán $u_{ij}^{(t)}$ theo công thức (16).

e) Xác định $K_i^{(t)}$ theo công thức (21).

f) Dự đoán giá trị $r_{ix}^{(t)}$ theo công thức (22).

2.2. Bán giám sát tập đánh giá sản phẩm và tập đặc trưng người dùng:

a) Xác định trọng số các đặc trưng nội dung người dùng $v_{qx}^{(t)}$ tại vòng lặp thứ t theo công thức (10).

b) Mở rộng ma trận đánh giá theo hồ sơ sản phẩm $r_{ix}^{(t)}$ theo công thức (11).

c) Xác định $S_x^{(t)}$ theo công thức (23).

d) Tính toán $p_{xy}^{(t)}$ theo công thức (27).

e) Xác định $K_x^{(t)}$ theo công thức (32).

f) Dự đoán giá trị $r_{ix}^{(t)}$ theo công thức (33).

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

until *Converges*.

Bước 3 *(sinh ra tư vấn):*

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

⟨Chọn k sản phẩm x đầu tiên tư vấn cho người dùng i⟩;

end

IV. ĐÁNH GIÁ THỰC NGHIỆM

Để đánh giá hiệu quả của các phương pháp tư vấn kết hợp đề xuất, chúng tôi tiến hành thử nghiệm trên bộ dữ

liệu thực về phim [14]. Phương pháp trình bày ở trên được đánh giá và so sánh với các phương pháp khác theo thủ tục mô tả dưới đây.

1. Dữ liệu thử nghiệm

Thuật toán học bán giám sát cho lọc kết hợp được thử nghiệm trên bộ dữ liệu MovieLens của nhóm nghiên cứu GroupLens thuộc trường đại học Minnesota [14]. Tập dữ liệu MovieLens có ba lựa chọn với kích thước khác nhau lần lượt là: MovieLens 100 KB, MovieLens 1 MB và MovieLens 10 MB. Trong đó, tập dữ liệu MovieLens 100 KB là tập con của tập MovieLens 1 MB. Tập dữ liệu MovieLens 1 MB cung cấp đầy đủ tập đặc trưng sản phẩm và người dùng kèm theo tập đánh giá người dùng. Tập dữ liệu MovieLens 10 M tuy lớn nhưng không cung cấp tập đặc trưng người dùng và tập đặc trưng sản phẩm. Chính vì vậy, chúng tôi sử dụng tập dữ liêu MovieLens 1 M để tiến hành thử nghiệm cho phương pháp đề xuất.

Tập dữ liệu MovieLens gồm 1MB đánh giá của 6040 người dùng cho 3952 phim. Giá trị đánh giá được thực hiện từ 1 đến 5. Mức độ thưa thớt dữ liệu đánh giá là 99.1%. Dữ liệu cụ thể được cung cấp trong các tệp tin sau [14]:

- u.data: Tệp tin lưu trữ đầy đủ 1 MB 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: Tệp tin lưu số lượng người dùng, số lượng sản phẩm, số lượng xếp hạng của tập dữ liệu.
- u.item: Tệp tin lưu thông tin về phim.
- u.genre: Tệp tin lưu danh sách 19 thể loại phim khác nhau. Đây là tập đặc trưng nội dung sản phẩm được dùng trong thử nghiệm phương pháp đề xuất. Ngoài ra, ứng với mỗi phim chúng tôi tách trong IMDB (Internet Movie Database) [15] để lấy tập đặc trưng nước sản xuất, hãng phim, đạo diễn, diễn viên chính để làm tập đặc trưng phim.
- u.user: Tệp tin 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: Tệp tin lưu danh sách các nghề nghiệp. Đây là tập đặc trưng nội dung người dùng được dùng trong thử nghiệm phương pháp đề xuất.

2. Phương pháp thử nghiệm

Trước tiên, toàn bộ dữ liệu thử nghiệm được chia thành hai phần, một phần U_{tr} được sử dụng làm dữ liệu huấn luyện, phần còn lại U_{te} được sử dụng để kiểm tra. Tập U_{tr} chứa 80% đánh giá và tập U_{te} chứa 20% đánh giá. Dữ liệu huấn luyện được sử dụng để xây dựng mô hình theo thuật toán mô tả ở trên. Với mỗi người dùng i thuộc tập dữ liệu kiểm tra, các đánh giá (đã có) của người dùng được chia làm hai phần O_i và P_i. O_i được coi là đã biết, trong khi đó P_i là đánh giá cần dự đoán từ dữ liệu huấn luyện và O_i [7, 8].

Sai số dự đoán MAE_u với mỗi khách hàng u thuộc tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số tuyệt đối giữa giá trị dự đoán và giá trị thực đối với tất cả mặt hàng thuộc tập P_u,

$$MAE_u = \frac{1}{|P_u|} \sum_{y \in P_u} \left| \hat{r}_{uy} - r_{uy} \right|. \tag{34}$$

Sai số dự đoán trên toàn tập dữ liệu kiểm tra, MAE, được tính bằng trung bình cộng sai số dự đoán cho mỗi khách hàng thuộc U_{te},

$$MAE = \frac{\sum_{u \in U_{\text{te}}} MAE_u}{|U_{\text{te}}|}. \tag{35}$$

Giá trị MAE nhỏ thì phương pháp dự đoán có độ chính xác cao [2, 7].

3. So sánh và đánh giá

Phương pháp học bán giám sát đề xuất trong mục 3 (ký hiệu là Semi-Learning) được thử nghiệm và so sánh với những phương pháp sau:

- Phương pháp tư vấn cộng tác dựa vào người dùng sử dụng độ tương quan Pearson (ký hiệu là CF-UserBased) [3, 9].
- Phương pháp tư vấn cộng tác dựa vào sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CF-ItemBased) [3, 10].
- Phương pháp tư vấn nội dung dựa vào hồ sơ người dùng sử dụng độ tương quan Pearson (ký hiệu là CBF-UserBased) [4].
- Phương pháp tư vấn nội dung dựa vào hồ sơ sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CBF-ItemBased) [5].
- 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 (16).
- 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 (27).

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,

Bảng VIII
GIÁ TRỊ *MAE* CỦA CÁC PHƯƠNG PHÁP

Phương pháp	Số lượng đánh giá biết trước trong tập kiểm tra			
	5	10	15	20
CBF-USERBASED	0,865	0,859	0,855	0,835
CBF-ITEMBASED	0,894	0,883	0,875	0,845
CF-USERBASED	0,824	0,817	0,821	0,813
CF-ITEMBASED	0,846	0,841	0,836	0,815
HYBRID-USERBASED	**0,793**	**0,792**	**0,791**	**0,702**
HYBRID-ITEMBASED	**0,798**	**0,788**	**0,782**	**0,695**
SEMI-LEARNING	**0,672**	**0,629**	**0,617**	**0,585**

test2.inp, test3.inp, test4.inp). Đối với mỗi tập dữ liệu kiểm tra, chúng tôi 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 dữ liệu kiểm tra (test1.inp, test2.inp, test3.inp, test4.inp) theo thứ tự để xác định xác định w_{is}, v_{qx} theo công thức (6), (10). 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 (12), (16), (21), và S_x, p_{xy}, K_x theo công thức (23), (27), (32). Giá trị *MAE* trong Bảng VIII được lấy trung bình của 10 lần thử nghiệm ngẫu nhiên. Giá trị *MAE* nhỏ chứng tỏ phương pháp có kết quả dự đoán tốt [2, 7, 12].

Kết quả trong Bảng VIII cho thấy phương pháp tư vấn nội dung dựa vào hồ sơ người dùng và hồ sơ sản phẩm cho lại giá trị *MAE* 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* 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. Kết quả này hoàn toàn phù hợp với những nghiên cứu trước đây [1–3].

Phương pháp Hybrid-UserBased cho lại giá trị *MAE* thấp hơn nhiều so với phương pháp CBF-UserBased và CF-UserBased. Cụ thể ứng với số lượng đánh giá biết trước trong tập kiểm tra là 5, 10, 15, 20 thì phương pháp CBF-UserBased và CF-UserBased cho lại giá trị *MAE* lần lượt là 0,865; 0,859; 0,855; 0,835 và 0,824; 0,817; 0,821; 0,813 so với 0,793; 0,792; 0,791; 0,702 của phương pháp Hybrid-UserBased. Phương pháp Hybrid-ItemBased cũng cho lại giá trị *MAE* thấp hơn so với phương pháp CBF-ItemBased và CF-ItemBased. Với số lượng đánh giá biết trước trong tập kiểm tra là 5, 10, 15, 20 thì phương pháp CBF-ItemBased và CF-ItemBased cho lại giá trị *MAE* lần lượt là 0,894; 0,833; 0,875; 0,845 và 0,846; 0,841; 0,836; 0,815 so với 0,798; 0,788; 0,782; 0,695 của phương pháp Hybrid-ItemBased. Điều này chỉ có thể lý giải phương pháp tính toán mức độ tương tự giữa các cặp người dùng trên 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 trên 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 Semi-Learning cho lại giá trị *MAE* thấp nhất ở tất cả các mức độ thưa thớt dữ liệu khác nhau. Đối với tập dữ liệu kiểm tra chỉ có 5 đánh giá biết trước, phương pháp Hybrid-UserBased và Hybrid-ItemBased cho lại giá trị *MAE* lần lượt là 0,793; 0,798 so với 0,672 của phương pháp Semi-Learning. Với tập dữ liệu kiểm tra chỉ có 10 đánh giá biết trước, phương pháp Hybrid-UserBased và Hybrid-ItemBased cho lại giá trị *MAE* lần lượt là 0,792; 0,788 so với 0,629 của phương pháp Semi-Learning. Với tập dữ liệu kiểm tra chỉ có 15 đánh giá biết trước, phương pháp Hybrid-UserBased và Hybrid-ItemBased cho lại giá trị *MAE* lần lượt là 0,791; 0,782 so với 0,617 của phương pháp Semi-Learning. Đặc biệt, với tập dữ liệu kiểm tra có 20 đánh giá biết trước, phương pháp cho lại giá trị *MAE* là 0,585. 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 chuyển giao kết quả dự đoán giữa quá trình bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm và tập đánh giá sản phẩm cùng tập đặc trưng người dùng đã hạn chế hiệu quả vấn đề dữ liệu thưa của các phương pháp lọc.

V. KẾT LUẬN

Bài báo đã đề xuất một mô hình hợp nhất giữa lọc cộng tác và lọc theo nội dung bằng phương pháp học bán giám sát. Phương pháp được tiến hành bằng cách hợp nhất biểu diễn giá trị các đặc trưng sản phẩm vào lọc cộng tác để thống nhất các phương pháp dự đoán dựa vào người dùng. Sau đó, xây dựng phương pháp hợp nhất biểu diễn giá trị các đặc trưng người dùng vào lọc cộng tác để thống nhất các phương pháp dự đoán dựa vào sản phẩm. Cuối cùng, chúng

tôi xây dựng phương pháp học bán giám sát để chuyển giao kết quả dự đoán giữa hai phương pháp dự đoán theo người dùng và dự đoán theo sản phẩm.

Để phát huy ưu điểm và hạn chế nhược điểm của các phương pháp lọc, chúng tôi đề xuất xây dựng hai kiểu bán giám sát: bán giám sát trên tập đánh giá người dùng cùng tập đặc trưng sản phẩm và bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng. Bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm được tiến hành bằng cách xây dựng tập không thưa đối với mỗi người dùng. Bán giám sát tập đánh giá sản phẩm cùng tập đặc trưng người dùng được tiến hành bằng cách xác định tập không thưa đối với mỗi sản phẩm. Dựa trên các tập không thưa đối với mỗi người dùng và sản phẩm, chúng tôi đã hạn chế được quá trình tính toán mức độ tương tự giữa các cặp người dùng, tập láng giềng của của người dùng và sản phẩm để xác định các kết quả dự đoán chắc chắn. Trên cơ sở của hai quá trình bán giám sát đã được xây dựng, chúng tôi đề xuất xây dựng thuật toán học bán giám sát để chuyển giao kết quả dự đoán giữa các quá trình bán giám sát. 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 trong trường hợp dữ liệu thưa.

TÀI LIỆU THAM KHẢO

[1] M. D. Ekstrand, J. T. Riedl, J. A. Konstan *et al.*, "Collaborative filtering recommender systems," *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011.

[2] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[3] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009, pp. 1–20, 2009.

[4] T. Miranda, M. Claypool, A. Gokhale, T. Mir, P. Murnikov, D. Netes, and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," in *In Proceedings of ACM SIGIR Workshop on Recommender Systems*, 1999.

[5] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp. 393–408, 1999.

[6] A. Gunawardana and C. Meek, "A unified approach to building hybrid recommender systems," in *Proceedings of the third ACM Conference on Recommender Systems*. ACM, 2009, pp. 117–124.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[8] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *Journal of Machine Learning Research*, vol. 10, no. Dec, pp. 2935–2962, 2009.

[9] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.

[10] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web*. ACM, 2001, pp. 285–295.

[11] R. Burke, F. Vahedian, and B. Mobasher, "Hybrid recommendation in heterogeneous networks," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2014, pp. 49–60.

[12] S. Raghavan, S. Gunasekar, and J. Ghosh, "Review quality aware collaborative filtering," in *Proceedings of the sixth ACM Conference on Recommender systems*. ACM, 2012, pp. 123–130.

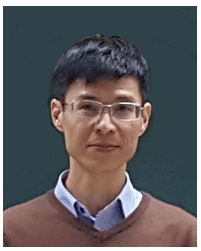
[13] J. Wang, A. P. De Vries, and M. J. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 501–508.

[14] http://www.grouplens.org/.

[15] http://www.imdb.com/.

Đỗ Thị Liên tốt nghiệp Đại học và nhận bằng Thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông vào các năm 2010 và 2013. Hiện nay, tác giả là giảng viên tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu chính của tác giả là học máy ứng dụng trong lọc thông tin và phát triển ứng dụng đa phương tiện.

Nguyễn Duy Phương tốt nghiệp Đại học và nhận bằng Thạc sĩ tại Trường Đại học Tổng hợp Hà Nội vào các năm 1988 và 1997. Năm 2010, ông bảo vệ luận án Tiến sĩ tại Đại học Quốc gia Hà Nội. Hiện nay, ông là Phó Trưởng khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu chính của ông là học máy ứng dụng trong lọc thông tin.



Từ Minh Phương tốt nghiệp Trường Đại học Bách khoa Taskent năm 1993 và bảo vệ Tiến sĩ tại Viện Hàn lâm Khoa học Uzbekistant, Taskent năm 1995. Hiện nay, ông là Phó Giáo sư, Trưởng Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu chính của ông là trí tuệ nhân tạo, học máy, tin sinh học.