*Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016*
*DOI: 10.15625/vap.2016.00052*

# MỘT PHƯƠNG PHÁP HỌC BÁN GIÁM SÁT CHO LỌC KẾT HỢP

**[1]Đỗ Thị Liên, [1]Nguyễn Duy Phương**

[1]Học viện Công nghệ Bưu chính Viễn thông

*liendt@ptit.edu.vn,phuongnd@ptit.edu.vn*

***TÓM TẮT***— *Hệ tư vấn (recommender systems) 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 (collaborative filtering) và lọc nội dung (content-based filtering). Lọc nội dung thực hiện hiệu quả trên các dạng thông tin văn bản nhưng gặp khó khăn trong trích chọn đặc trưng đối với 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 mô hình lọc kết hợp giữa lọc cộng tác và lọc nội dung bằng phương pháp học bán giám sát. Mô hình 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 phù hợp cao đối với người dùng này. Bán giám sát tập láng giềng theo sản phẩm cùng tập đặc trưng người dùng cho phép ta phát hiện ra những người dùngmới có khả năng phù hợp cao đối với sản phẩm này. Hai thủ tục bán giám sát thực hiện đồng thời và bổ sung qua lại cho nhau các giá trị dự đoán chắc chắn để nâng cao kết quả tư vấn. Kết quả thử nghiệm trên các bộ dữ liệu thực về phim cho thấy phương pháp đề xuất tận dụng hiệu quả ưu điểm và hạn chế đáng kể nhược điểm của mỗi phương pháp lọc.*

***Từ khóa***— *Tư vấn cộng tác, tư vấn theo nội dung, hệ tư vấn lai, tư vấn bằng phương pháp học có giám sát, tư vấn bằng phương pháp học không giám sát, tư vấn bằng phương pháp học bán giám sát.*

## I. GIỚI THIỆU VẤN ĐỀ

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…) phù hợp cho mỗi người dùng. Trên thực tế, hệ tư vấn không chỉ hướng đến vấn đề giảm tải thông tin cho mỗi người dùng mà nó còn là yếu tố quyết định đến thành công của các hệ thống thương mại điện tử. Bài toán tư vấn tổng quát có thể được phát biểu như sau.

Cho tập hợp hữu hạn gồm $N$ người dùng $U = \{u_1, u_2,\ldots, u_N\}$, $P = \{p_1, p_2,.., p_M\}$ là tập hữu hạn 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. Mối quan hệ giữa tập người dùng $U$ và tập sản phẩm $P$ được biểu diễn thông qua ma trận đánh giá $R=\{ r_{ix}: i = 1, 2, ..N; x = 1, 2,..M \}$. Giá trị $r_{ix}$ thể hiện đánh giá của người dùng $u_i \in U$ cho một số sản phẩm $p_x \in P$. Thông thường giá trị $r_{ix}$nhận một giá trị thuộc miền $F = \{ 1, 2,.., g\}$ được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị $r_{ix} = \phi$ được hiểu người dùng $u_i$ chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm $p_x$. Ma trận đánh giá của các hệ thống tư vấn thực tế thường rất thưa. Mật độ các giá trị $r_{ix} \neq 0$ nhỏ hơn 1%, hầu hết các giá trị $r_{ix}$ còn lại là $\phi$ [1, 17]. Ma trận $R$ chính là đầu vào của các hệ thống tư vấn cộng tác [3, 18]. Để thuận tiện trong trình bày, ta viết $p_x \in P$ ngắn gọn là$x \in P$; và $u_i \in U$ là $i \in U$. Các ký tự $i$, $j$ luôn được dùng để chỉ tập người dùng trong các mục tiếp theo của bài báo.

Mỗi sản phẩm $x \in P$ được biểu diễn thông qua $|C|$ đặc trưng nội dung $C = \{ c_1, c_2,.., c_{|C|}\}$. Các đặc trưng $c_s \in C$ nhận được từ các phương pháp trích chọn đặc trưng (feature selection) 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…*$\}$. Gọi $w_x = \{w_{x1}, w_{x2},.., w_{x|C|} \}$ là vector trọng số các giá trị đặc trưng nội dung sản phẩm $c_s \in C$ đối với mỗi sản phẩm $x \in P$. Khi đó, ma trận trọng số $W =\{w_{xs}: x =1, 2, .., M; s =1, 2, .., |C|\}$ chính là đầu vào của các hệ thống tư vấn theo nội dung sản phẩm [2, 3, 7]. Để thuận tiện trong trình bày, ta viết $c_s \in C$ ngắn gọn là $s \in C$.Ký tự $s$ luôn được dùng để chỉ tập đặc trưng nội dung sản phẩm trong các mục tiếp theo của bài báo.

Mỗi người dùng $i \in U$ được biểu diễn thông qua $|T|$ đặc trưng nội dung $T = \{t_1, t_2,.., t_{|T|}\}$. Các đặc trưng $t_q \in T$ 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,…*$\}$. Gọi $v_i = \{v_{i1}, v_{i2},.., v_{i|T|}\}$ là vector trọng số biểu diễn các giá trị đặc trưng nội dung $t_q \in T$ đối với mỗi người dùng $i \in U$. Khi đó, ma trận trọng số $V =\{v_{iq}: i = 1, 2, .., N;q = 1, 2, .., |T| \}$ chính là đầu vào của các hệ thống tư vấn theo nội dung thông tin người dùng [3, 6]. Để thuận tiện trong trình bày, ta viết $t_q \in T$ ngắn gọn là $q \in T$. Ký tự $q$ luôn được dùng để chỉ tập đặc trưng nội dung người dùng trong các mục tiếp theo của bài báo.

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 $i \in U$ đã đánh giá sản phẩm $x \in P$. Với một người dùng cần được tư vấn $i \in U$ (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 $x \in (P \backslash P_i)$ phù hợp nhất đối với người dùng $i$.

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, 2]. 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 $W=\{w_{xs}\}$ hoặc ma trận trọng số các đặc trưng nội dung người dùng $V=\{v_{iq}\}$ [6, 7]. Các đặc trưng nội dung $s \in C$ được xây dựng từ các kỹ thuật truy vấn thông tin. Trọng số của mỗi đặc trưng nội dung $w_{xs}$ thường được ước lượng bằng kỹ thuật tf-idf [3, 17, 18]. 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…). Một số đặc trưng nội dung không quan trọng đối với sản phẩm vẫn được ước lượng với trọng số cao trong khi một số đặc trưng nội dung quan trọng bị bỏ qua trong quá trình trích chọn đặc trưng [2, 17]. Một người dùng mới tham gia hệ thống sẽ có hồ sơ sử dụng sản phẩm là $\{\phi\}$. Khi đó, hệ thống sẽ không thể gợi ý được các sản phẩm phù hợp với người dùng này [2, 17].

Hệ tư vấn cộng tác xây dựng phương pháp dự đoán dựa trên ma trận đánh giá $R=\{r_{ix}\}$ [8,9, 12, 13]. 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…). 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 [1, 15]. Thách thức lớn nhất của lọc cộng tác là vấn đề dữ liệu thưa, người dùng mới và sản phẩm mới. Vấn đề dữ liệu thưa xảy ra khi số lượng giá trị đánh giá biết trước ít hơn rất nhiều số lượng đánh giá chưa biết [1, 18]. Một người dùng mới tham gia hệ thống sẽ có hồ sơ sử dụng sản phẩm là $\{\phi\}$, khi đó ta không thể gợi ý các sản phẩm phù hợp đối với người dùng này [18]. Một sản phẩm mới chưa được bất kỳ người dùng nào đánh giá thì hệ thống cũng không có cơ sở gợi ý sản phẩm này cho bất kỳ người dùng nào [17, 18].

Hệ tư vấn lai xây dựng phương pháp dự đoán dựa trên cả ba ma trận $R$, $W$,$V$[2, 5, 11, 14]. Giá trị $r_{ix}$ phản ánh quan điểm của người dùng $i \in U$ đối với các sản phẩm $x \in P$, $w_{xs}$ phản ánh mức độ quan trọng của đặc trưng $s \in C$ đối với sản phẩm$x \in P$, $v_{iq}$ phản ánh mức độ quan trọng của đặc trưng $q \in T$ đối với người dùng $i \in U$. Hệ tư vấn lai được tiếp cận theo bốn su 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 lai 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.

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 hạn 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.

Để trọng tâm vào những đóng góp mới của bài báo, Mục tiếp theo chúng tôi trình bày phương pháp ước lượng trọng số các đặc trưng nội dung người dùng và sản phẩm của lọc nội dung. Mục 3 trình bày phương pháp học bán giám sát dựa vào đánh giá người dùng, đặc trưng sản phẩm và đặc trưng người dùng. Mục 4 trình bày phương pháp thử nghiệm và đánh giá. Mục cuối dùng 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 lai.

Không hạn chế tính tổng quát của bài toán phát biểu trong Mục 1, 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 (1). Mỗi sản phẩm $x \in P$ được biểu diễn thông qua $|C|$ đặc trưng nội dung $C = \{c_1, c_2,..,c_{|C|}\}$ được xác định theo công thức (2). Mỗi người dùng $i \in U$ được biểu diễn thông qua $|T|$ đặc trưng nội dung $T = \{t_1, t_2,.., t_{|T|}\}$ được xác định theo công thức (3).

$$r_{ix} = \begin{cases} v & \text{nếu người dùng } i \text{ đánh giá sản phẩm } x \text{ ở mức độ } v\ (v \in F) \\ 0 & \text{nếu người dùng } i \text{ chưa đánh giá hoặc chưa biết đến sản phẩm } x \end{cases} \quad (1)$$

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

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

Ví dụ với hệ gồm *3* người dùng $U = \{u_1, u_2, u_3\}$, *4* sản phẩm $P = \{p_1, p_2, p_3, p_4\}$. Trong đó, ma trận đánh giá $R$ được cho trong Bảng 1; Ma trận đặc trưng nội dung sản phẩm $C$ được cho trong Bảng 2; Ma trận đặc trưng nội dung người dùng $T$ được cho trong Bảng 3. Hệ tư vấn cộng tác được xây dựng dựa trên ma trận đánh giá $R$ [3, 13, 14]. Hệ tư

**Bảng 1**. Ma trận đánh giá R

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

**Bảng 2.** Ma trận đặc trưng sản phẩm C

| | | | |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |

**Bảng 3.** Ma trận đặc trưng người dùng T

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

vấn nội dung được xây dựng dựa trên ma trận các đặc trưng nội dung $C$ và $T$ [2, 4, 5, 6]. Hệ tư vấn lai xây dựng dựa trên ma cả ba ma trận $R$,$C$ và $T$ [2, 5, 16].

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

Phương pháp tư vấn theo nội dung thực hiện dự đoán các sản phẩm có nội dung thông tin hay mô tả hàng hóa tương tự với những sản phẩm mà người dùng đã từng sử dụng hoặc truy nhập trong quá khứ. Chất lượng của các phương pháp tư vấn theo nội dung phụ thuộc vào phương pháp trích chọn đặc trưng để biểu diễn vector đặc trưng nội dung sản phẩm và vector hồ sơ sử dụng sản phẩm của người dùng. Hạn chế lớn nhất của phương pháp trích chọn đặc trưng hiện nay là nhiều đặc trưng không quan trọng nhưng vẫn tham gia vào việc xác định mức độ tương tự giữa vector hồ sơ người dùng và vector đặc trưng sản phẩm [2]. Để hạn chế điều này, chúng tôi đề xuất phương pháp xây dựng hồ sơ sử dụng các đặc trưng sản phẩm của người dùng thông qua đánh giá tự nhiên của người dùng đối với sản phẩm. Phương pháp được tiến hành như dưới đây.

Để xây dựng được hồ sơ sử dụng các đặc trưng sản phẩm của người dùng ta cần thực hiện hai nhiệm vụ: xác định được tập các 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, 5, 7]. Gọi $P_i \subseteq P$ được xác định theo công thức (4) là tập sản phẩm người dùng $i \in U$ đã đánh giá các sản phẩm $x \in P$. 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 $i \in U$.

$$P_i = \{x \in P \mid r_{ix} \neq 0 \ \ (i \in U)\} \tag{4}$$

Gọi $Item(i, s)$ là tập các sản phẩm $x \in P_i$ chứa đựng đặc trưng $s \in C$ được xác định theo công thức (5). Khi đó, $|Item(i , s)|$ chính là số lần người dùng $i \in U$ sử dụng các sản phẩm $x \in P$ chứa đựng đặc trưng $s \in C$ trong quá khứ.

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

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 thường được sử dụng trong xây dựng hồ sơ người dùng là kỹ thuật tf-idf [7]. 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 [8, 9], 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 được số lượng trung bình của tất cả người dùng $i \in U$ 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 $i \in U$ 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.

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

Giá trị $w_{is}$ được ước lượng theo (6) phản ánh quan điểm của người dùng $i \in U$ đối với các đặc trưng nội dung sản phẩm $s \in C$ cũng chính là hồ sơ người dùng $i \in U$ đã sử dụng các đặc trưng nội dung $s \in C$ trong quá khứ. Dễ dàng nhận thấy $w_{is} \in F$, trong đó $F = \{ 1, 2, .., g\}$. Chính vì vậy, ta có thể xem mỗi đặc trưng nội dung sản phẩm đóng vai trò như một sản phẩm phụ bổ sung vào tập sản phẩm. Dựa trên nhận xét này, 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 (7). Trong đó, $x = s$ ($s \in C$) đóng vai trò như một sản phẩm phụ bổ để mở rộng ma trận đánh giá về phía sản phẩm.

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

Ví dụ với hệ có ma trận đánh giá theo Bảng 1, ma trận đặc trưng sản phẩm theo Bảng 2, ma trận đặc trưng người dùng theo Bảng 3, chọn θ = 2, khi đó ta sẽ tính toán được tập hồ sơ người dùng $\{w_{is} : i \in U, s \in C\}$ trong Bảng 4 và ma trận đánh giá mở rộng theo (7) trong Bảng 5.

**Bảng 4**. 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 5**. 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 |

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 theo người dùng đều có thể dễ dàng triển khai trên ma trận đánh giá mở rộng theo hồ sơ người dùng. Phương pháp tư vấn cộng tác theo người dùng được xây dựng dựa vào tập đánh giá người dùng [9]. Phương pháp tư vấn theo nội dung sản phẩm được thực hiện dựa trên hồ sơ người dùng [7]. Phương pháp tư vấn lai được thực hiện dựa vào tập đánh giá người dùng và hồ sơ người dùng [14]. 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.

### 2.2. 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 ta cần thực hiện hai nhiệm vụ: xác định được tập người dùng đã từng sử dụng sản phẩm quá khứ và ước lượng trọng số mỗi đặc trưng nội dung người dùng trong hồ sơ sản phẩm [6, 14]. Gọi $U_x \subseteq U$ được xác định theo công thức (8) là tập người dùng $i \in U$ 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 $x \in P$.

$$U_x = \{i \in U \mid r_{ix} \neq 0\ \ (x \in P)\} \quad (8)$$

Gọi *User* $(x, q)$ là tập người dùng $i \in U_x$ có đặc trưng $q \in T$ được xác định theo công thức (9). Khi đó, $|User(x, q)|$ chính là số lần sản phẩm $x \in P$ được tập người dùng $i \in U$ có đặc trưng nội dung $q \in T$ sử dụng trong quá khứ.

$$User(x, q) = \{i \in U_x \mid t_{iq} \neq 0\ \ (x \in P, q \in T\} \quad (9)$$

Dựa trên $U_x$ và *User*$(x, q)$ các phương pháp tư vấn theo nội dung người dùng ước lượng được trọng số $t_{xq}$ phản ánh mức độ quan trọng của đặc trưng nội dung $q$ đối với sản phẩm $x$. 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 $v_{xq}$ 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 $v_{xq}$ được tính bằng tổng của tất cả các giá trị đánh giá chia cho θ. Trong thử nghiệm, chúng tôi tính toán được số lượng trung bình của tất cả sản phẩm $x \in P$ được đánh giá bởi người dùng $i \in U$, sau đó chọn θ tương đương với 2/3 số lượng người dùng $i \in U$ chứa đựng đặc trưng $q \in T$ đã sử dụng sản phẩm $x \in P$. Bằng cách này ta có thể hạn chế được một số đặc trưng nội dung người dùng ít quan tâm đến sản phẩm nhưng vẫn được đánh giá với trọng số cao.

$$v_{qx} = \begin{cases} \frac{1}{|User(x,q)|} \sum_{i \in User(x,q)} r_{ix} & nếu\ \ |User(x, q)| \geq \theta \\ \frac{1}{\theta} \sum_{i \in User(x,q)} r_{ix} & nếu\ \ |User(x, q)| < \theta \end{cases} \quad (10)$$

Giá trị $v_{qx}$ được ước lượng theo (10) biểu diễn hồ sơ sản phẩm $x \in P$ đã được tập những người dùng $i \in U$ chứa đựng đặc trưng $q \in T$ sử dụng. Dễ dàng nhận thấy $v_{xq} \in F$, trong đó $F = \{1, 2, .., g\}$. Chính vì lý do này, ta có thể xem mỗi đặc trưng nội dung người dùng đóng vai trò như một người dùng phụ bổ sung vào tập người dùng. Dựa trên nhận xét này, 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 (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.

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

Ví dụ với hệ có ma trận đánh giá theo Bảng 1, ma trận đặc trưng người dùng theo Bảng 3, chọn θ = 2, khi đó ta sẽ tính toán được tập hồ sơ sản phẩm $\{v_{qx}: x \in P, q \in T\}$ trong Bảng 6 và ma trận đánh giá mở rộng về phía người dùng theo (11) trong Bảng 7.

**Bảng 6**. 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 7**. Ma trận đánh giá mở rộng $r_{ix}$ theo hồ sơ sản phẩm

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

Hệ tư vấn được xác định theo (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 theo sản phẩm đều có thể dễ dàng triển khai trên ma trận đánh giá mở rộng theo hồ sơ sản phẩm. Phương pháp tư vấn cộng tác theo sản phẩm được xây dựng dựa vào tập đánh giá sản phẩm [10, 13]. Phương pháp tư vấn theo nội dung người dùng được thực hiện dựa trên hồ sơ sản phẩm [6, 10]. Phương pháp tư vấn lai được thực hiện dựa vào tập đánh giá sản phẩm và hồ sơ sản phẩm [6, 14]. 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 (7), (11) đều gặp phải vấn đề dữ liệu thưa [1, 12, 15]. 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 với tập đặc trưng sản phẩm và bán giám sát tập đánh giá sản phẩm cùng với tập đặc trưng người dùng. Bán giám sát tập đánh giá người dùng cùng tập đặc trưng sản phẩm cho phép ta dự đoán được những sản phẩm mới có khả năng cao phù hợp cho mỗi người dùng. Những sản phẩm mới được dự đoán được sẽ được 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 các giá trị dự đoán chắc chắn cho nhau để nâng cao chất lượng tư vấn.

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

Hệ tư vấn lai được xác định theo (7) cho phép ta dễ dàng triển khai các phương pháp lọc cộng tác dựa vào người dùng [9, 14, 15]. Phương pháp được tiến hành thông qua 4 bước: tính toán mức độ tương tự giữa các cặp người dùng, xác định tập láng giềng cho người dùng cần tư vấn, dự đoán quan điểm của người dùng đối với các sản phẩm mới, và tư vấn top k sản phẩm có giá trị dự đoán cao nhất cho người dùng [9, 15]. Do tính chất thưa thớt của ma trận đánh giá làm cho việc xác định mức độ tương tự giữa các cặp người dùng kém chính xác. Điều này sẽ ảnh hưởng trực tiếp đến việc xác định tập láng giềng và kết quả dự đoán các sản phẩm mới cho người dùng cần được tư vấn [14]. Để khắc phục điều này, với mỗi người dùng $i \in U$ chúng tôi xây dựng tập $S_i$ được định nghĩa theo công thức (12) để giám sát việc tính toán mức độ tương tự giữa các cặp người dùng. Trong đó, $P_i$ được xác định theo công thức (4), $C_i$ được xác định theo công thức (13).

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

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

$S_i$ được xác định theo (12) là tập người dùng $j \in U$ có số lượng đánh giá giao nhau với người dùng $i$ ít nhất là $\theta_1$ sản phẩm và số lượng các đặc trưng sản phẩm giao nhau ít nhất là $\theta_2$. Hai hằng số nguyên dương $\theta_1$ và $\theta_2$ 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, 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).

$$a_{ij} = \begin{cases} 0 & nếu\ j \notin S_i \\ \dfrac{\sum_{x \in P_i \cap P_j}(r_{ix} - \bar{r_i})(r_{jx} - \bar{r_j})}{\sqrt{\sum_{x \in P_i \cap P_j}(r_{ix} - \bar{r_i})^2}\sqrt{\sum_{x \in P_i \cap P_j}(r_{jx} - \bar{r_j})^2}} & nếu\ j \in S_i \end{cases} \quad (14)$$

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

$$u_{ij} = \begin{cases} \dfrac{\sum_{x \in H_i \cap H_j}(r_{ix} - \overline{\overline{r}_i})(r_{jx} - \overline{\overline{r}_j})}{\sqrt{\sum_{x \in H_i \cap H_j}(r_{ix} - \overline{\overline{r}_i})^2}\sqrt{\sum_{x \in H_i \cap H_j}(r_{ix} - \overline{\overline{r}_i})^2}} & nếu\ j \in S_i\ và\ a_{ij} \geq \alpha và\ b_{ij} \geq \alpha \\ 0 & trong\ các\ trường\ hợp\ khác \end{cases} \quad (16)$$

Trong đó, $P_i$ được xác định theo công thức (4), $C_i$ được xác định theo công thức (13); $H_i$, $\overline{r_i}$ , $\ddot{r}_i$, $\overline{\overline{r_i}}$ được xác định theo công thức (17), (18), (19), (20), theo thứ tự.

$$H_i = P_i \cup C_i \quad (17)$$

$$\overline{r_i} = \frac{1}{|P_i \cap P_j|}\sum_{x \in P_i \cap P_j} r_{ix} \quad (18)$$

$$\ddot{r}_i = \frac{1}{|C_i \cap C_j|}\sum_{s \in C_i \cap C_j} r_{is} \quad (19)$$

$$\overline{\overline{r_i}} = \frac{1}{|H_i \cap H_j|}\sum_{x \in H_i \cap H_j} r_{ix} \quad (20)$$

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 $U \cup C$. 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 $\alpha$ nào đó. Ngưỡng $\alpha$ 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 (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 (22) [9, 15, 16].

$$K_i = \{j \in S_i:\ u_{ij} > \alpha\} \quad (21)$$

$$r_{ix} = \overline{\overline{r_i}} + \frac{\sum_{j \in K_i}(r_{jx} - \overline{\overline{r_j}})u_{ij}}{\sum_{j \in K_i}|u_{ij}|} \quad (22)$$

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.

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

Hệ tư vấn lai được xác định theo (19) cho phép ta dễ dàng triển khai các phương pháp lọc cộng tác dựa vào sản phẩm [10, 15]. Phương pháp được tiến hành thông qua 4 bước: tính toán mức độ tương tự giữa các cặp sản phẩm, xác định tập láng giềng cho sản phẩm cần tư vấn, dự đoán quan mức độ phù hợp của sản phẩm đối với mỗi người dùng, và tư vấn top k sản phẩm có giá trị dự đoán cao nhất cho người dùng [10]. Do tính chất thưa thớt của ma trận đánh giá làm cho việc xác định mức độ tương tự giữa các cặp sản phẩm kém chính xác. Điều này sẽ ảnh hưởng trực tiếp đến việc xác định tập láng giềng của sản phẩm và kết quả dự đoán mức độ phù hợp của người dùng đối với sản phẩm [1,10]. Để khắc phục điều này, với mỗi sản phẩm $x \in P$ chúng tôi xây dựng tập $S_x$ được định nghĩa theo công thức (23) để giám sát việc tính toán mức độ tương tự giữa các cặp sản phẩm. Trong đó, $U_x$ được xác định theo công thức (8), $T_x$ được xác định theo công thức (24).

$$S_x = \{y \in P: |U_x \cap U_y| \geq \gamma 1 và\ |T_x \cap T_y| \geq \gamma 2\} \quad (23)$$

$$T_x = \{q \in T: r_{qx} \neq 0\} \quad (24)$$

$S_x$ được xác định theo (23) là tập sản phẩm $y \in P$ có số lượng người dùng đánh giá với sản phẩm $x$ giao nhau ít nhất là $\gamma_1$ và số lượng các đặc trưng người dùng giao nhau ít nhất là $\gamma_2$. Hai hằng số nguyên dương $\gamma_1$ và $\gamma_2$ 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).

$$a_{xy} = \begin{cases} 0 & nếu\ y \notin S_x \\ \dfrac{\sum_{i \in U_x \cap U_y}(r_{ix} - \overline{r_x})(r_{iy} - \overline{r_y})}{\sqrt{\sum_{i \in U_x \cap U_y}(r_{ix} - \overline{r_x})^2}\sqrt{\sum_{i \in U_x \cap U_y}(r_{iy} - \overline{r_y})^2}} & nếu\ y \in S_x \end{cases} \quad (25)$$

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

$$p_{xy} = \begin{cases} \dfrac{\sum_{i \in H_x \cap H_y}(r_{ix} - \overline{\overline{r_x}})(r_{iy} - \overline{\overline{r_y}})}{\sqrt{\sum_{i \in H_x \cap H_y}(r_{ix} - \overline{\overline{r_x}})^2}\sqrt{\sum_{i \in H_x \cap H_y}(r_{iy} - \overline{\overline{r_y}})^2}} & nếu\ y \in S_x\ và\ a_{xy} \geq \alpha\ và\ b_{xy} \geq \alpha \\ 0 & trong\ các\ trường\ hợp\ khác \end{cases} \quad (27)$$

Trong đó, $U_x$ 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 theo công thức (28), (29), (30), (31), theo thứ tự .

$$H_x = U_x \cup T_x \quad (28)$$

$$\overline{r_x} = \frac{1}{|U_x \cap U_y|}\sum_{i \in U_x \cap U_y} r_{ix} \quad (29)$$

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

$$\overline{\overline{r_x}} = \frac{1}{|H_x \cap H_y|}\sum_{i \in H_x \cap H_y} r_{ix} \quad (31)$$

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 gười 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 $P \cup T$. 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 α=0.90 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 (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 (33)[10, 15, 16].

$$K_x = \{y \in S_x :\ p_{xy} > \alpha\} \quad (32)$$

$$r_{ix} = \frac{\sum_{y \in K_x} p_{xy} r_{iy}}{\sum_{y \in K_x} |p_{xy}|} \quad (33)$$

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 và 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.

### 3.3. 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 được mô tả chi tiết như trong Hình 1.

Thuật toán đề xuất ký hiệu là (Semi-Learning) thực hiện tuần tự thông qua ba bước: bước khởi tạo, bước lặp và tạo nên tư vấn. Tại bước khởi tạo *t=0*, ma trận ghi lại kết quả dự đoán được khởi tạo bằng chính ma trận đánh giá ban đầu của lọc cộng tác $R^{(0)} = \left\{r_{ij}^{(0)} = r_{ij} : i = 1, 2, \dots, N; j = 1, 2, \dots, M\right\}$. Tại bước lặp, quá trình bán giám sát theo đánh giá người dùng và tập đặc trưng sản phẩm được thực hiện tuần tự theo các bước (2.1.a), (2.1.b), (2.1.c), (2.1.d), (2.1.e), (2.1.f). Tại bước (2.1.a) ta xác định được giá trị $w_{is}^{(t)}$ phản ánh quan điểm của người dùng $i \in U$ đối với các đặc trưng sản phẩm $s \in C$ của vòng lặp thứ (t) theo công thức (6). Sử dụng $w_{is}^{(t)}$, tại bước (2.1.b) ta xây dựng được ma trận đánh giá mở rộng theo hồ sơ người dùng của vòng lặp thứ (t) theo công thức (7). Dựa vào kết quả của bước (2.1.b), tại bước (2.1.c) ta xác định được tập $S_i^{(t)}$là tập dữ liệu không thưa đối với người dùng $i \in U$ củavòng lặp thứ (t) theo công thức

(12). Sử dụng $S_i^{(t)}$, bước (2.1.d) ta xác định được $u_{ij}^{(t)}$ là mức độ tương tự giữa các cặp người dùng *i, j*$\in U$ trên cả tập đánh giá người dùng và tập đặc trưng sản phẩm của vòng lặp thứ (t) theo công thức (16). Sau khi tính toán được $u_{ij}^{(t)}$, tại bước (2.1.e) ta xác định được $K_i^{(t)}$ là tập láng giềng của người dùng *i*của vòng lặp thứ (*t*) theo công thức (21). Cuối cùng, tại bước (2.1.f) ta dự đoán được giá trị $r_{ix}^{(t)}$ phản ánh quan điểm của người dùng *i* đối với sản phẩm mới *x*$\in P$ của vòng lặp thứ (t) theo công thức (22). 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 sản phẩm cùng tập đặc trưng người dùng tại bước 2.2 của thuật toán.

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 $u_{ij}^{(t)} = u_{ij}^{(t-1)}$ *v*à $p_{xy}^{(t)} = p_{xy}^{(t-1)}$ vì

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

$$p_{xy}^{(t)} = p_{xy}^{(t-1)} \Leftrightarrow \begin{cases} S_x^{(t)} = S_x^{(t-1)} \\ K_x^{(t)} = K_x^{(t-1)} \end{cases}$$

Điều này có nghĩa, tại vòng lặp thứ (t) ta không bổ sung được bất kỳ giá trị $r_{ix}^{(t)}$ nào theo cả hai quá trình bán giám sát. 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*.

**Đầu vào:**

- *Ma trận đánh giá R ={$r_{ix}$: i=1, 2,.., N; x =1, 2, .., M} được xác định theo (1).*
- *Ma trận các đặc trưng nội dung sản phẩm C ={$c_{xs}$: x=1, 2,.., M; s =1, 2, .., |C|}được xác định theo (2).*
- *Ma trận các đặc trưng nội dung người dùng T ={$c_{iq}$: i=1, 2,.., N; q =1, 2, .., |T|}được xác định theo (3).*
- *Người dùng i*$\in U$ *là người dùng cần được tư vấn.*

**Đầu ra :***Ma trận dự đoán* $R = R^{(t)} = \left\{r_{ix}^{(t)} : i = 1, 2, .., N;\ x = 1, 2, .., M\right\}$.

**Các bước tiến hành:**

**Begin**

**Bước 1**( *Khởi tạo*):

*t*←*0;//khởi tạo số bước lặp ban đầu là 0*

$R^{(0)} = \left\{r_{ix}^{(0)} = r_{ix} : i = 1, 2, \dots, N; x = 1, 2, \dots, M\right\}$; *//Khởi tạo ma trận đánh giá ban đầu tại vòng lặp thứ 0.*

**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 tại vòng lặp thứ t theo công thức* (6):

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

b) *Mở rộng ma trận đánh giá theo hồ sơ người dùng bằng công thức* (7):

$$r_{ix}^{(t)} = \begin{cases} r_{ix}^{(t)} = r_{ix}^{(t)} & nếu\ x \in P \\ w_{is}^{(t)} & nếu\ s \in C\ (x = s) \end{cases}$$

c) *Xác định* $S_i^{(t)}$ *theo công thức* (12):$S_i^{(t)} = \left\{j \in U : |P_i^{(t)} \cap P_j^{(t)}| > \theta_1\ v\text{à}\ |C_i^{(t)} \cap C_j^{(t)}| > \theta_2\right\}$

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

$$u_{ij}^{(t)} = \begin{cases} \dfrac{\sum_{x \in H_i^{(t)} \cap H_j^{(t)}} \left(r_{ix}^{(t)} - \overline{r_i^{(t)}}\right)\left(r_{jx}^{(t)} - \overline{r_j^{(t)}}\right)}{\sqrt{\sum_{x \in H_i^{(t)} \cap H_j^{(t)}} \left(r_{ix}^{(t)} - \overline{r_i^{(t)}}\right)^2} \sqrt{\sum_{x \in H_i^{(t)} \cap H_j^{(t)}} \left(r_{jx}^{(t)} - \overline{r_j^{(t)}}\right)^2}} & nếu\ j \in S_i^{(t)}\ và\ a_{ij}^{(t)} \geq \propto\ và\ b_{ij}^{(t)} \geq \propto \\ 0 & trong\ các\ trường\ hợp\ khác \end{cases}$$

e) Xác định $K_i^{(t)}$ theo công thức (21):$K_i^{(t)} = \left\{j \in S_i^{(t)} : u_{ij}^{(t)} > \alpha\right\}$

f) Dự đoán giá trị $r_{ij}^{(t)}$ theo công thức (22):$r_{ix}^{(t)} = \overline{r_i^{(t)}} + \dfrac{\sum_{j \in K_i^{(t)}} \left(r_{jx}^{(t)} - \overline{r_j^{(t)}}\right) u_{ij}^{(t)}}{\sum_{j \in K_i^{(t)}} u_{ij}^{(t)}}$

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 tại vòng lặp thứ t theo công thức* (10):

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

b) *Mở rộng ma trận đánh giá theo hồ sơ sản phẩm bằng công thức* (11):

$$r_{ix}^{(t)} = \begin{cases} r_{ix}^{(t)} = r_{ix}^{(t)} & nếu\ i \in U \\ v_{qx}^{(t)} & nếu\ q \in T\ (i = q) \end{cases}$$

c) *Xác định* $S_x^{(t)}$ *theo công thức* (23):$S_x^{(t)} = \left\{y \in P : |U_x^{(t)} \cap U_y^{(t)}| > \theta_1\ và\ |T_x^{(t)} \cap T_y^{(t)}| > \theta_2\right\}$

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

$$p_{xy}^{(t)} = \begin{cases} 0 & nếu\ y \notin S_x^{(t)} \\ \dfrac{\sum_{i \in H_x^{(t)} \cap H_y^{(t)}} \left(r_{ix}^{(t)} - \overline{r_x^{(t)}}\right)\left(r_{iy}^{(t)} - \overline{r_y^{(t)}}\right)}{\sqrt{\sum_{i \in H_x^{(t)} \cap H_y^{(t)}} \left(r_{ix}^{(t)} - \overline{r_x^{(t)}}\right)^2} \sqrt{\sum_{i \in H_x^{(t)} \cap H_y^{(t)}} \left(r_{iy}^{(t)} - \overline{r_y^{(t)}}\right)^2}} & nếu\ i \in S_x^{(t)}\ và\ a_{xy}^{(t)} \geq \propto\ và\ b_{xy}^{(t)} \geq \propto \end{cases}$$

e) Xác định $K_x^{(t)}$ theo công thức (32):$K_x^{(t)} = \left\{x \in S_x^{(t)} : p_{xy}^{(t)} > \alpha\right\}$

f) Dự đoán giá trị $r_{ix}^{(t)}$ theo công thức (33):$r_{ix}^{(t)} = \dfrac{\sum_{y \in K_x^{(t)}} p_{xy}^{(t)} r_{iy}^{(t)}}{\sum_{y \in K_x^{(t)}} |p_{xy}^{(t)}|}$

2.3. Tăng bước lặp : *t←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 top k sản phẩm x đầu tiên tư vấn cho người dùng i*>;

**End**.

**Hình 1.** Thuật toán Semi-Learning

## 4. THỬ NGHIỆM VÀ ĐÁNH GIÁ

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

### 4.1. 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 [18]. Tập dữ liệu MovieLens có ba lựa chọn với kích thước khác nhau lần lượt là: MovieLens 100k, MovieLens 1M và MovieLens 10M. Trong đó, tập dữ liệu MovieLens 100KBlà tập con của tập MovieLens 1M. Tập đặc trưng sản phẩm và người dùng cũng được cung cấp đầy đủ kèm theo tập đánh giá người dùng. Tập dữ liệu MovieLens 10M tuy lớn nhưng không cung cấp tập đặc trưng người dùng và tập đặc trưng sản phẩm. Chính vì vậy, chúng tôi sử dụng tập dữ liêu MovieLens 1M để tiến hành thử nghiệm cho phương pháp đề xuất.

Tập dữ liệu MovieLens 1M 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 file sau [18]:

- u.data: lưu trữ đầy đủ 1MB đánh giá của 6040 người dùng cho 3952 phim. Mỗi người dùng đánh giá ít nhất 20 phim. Mỗi hàng đều có cùng cấu trúc: user id | item id | rating | timestamp.
- u.info: File lưu số lượng người dùng, số lượng sản phẩm, số lượng xếp hạng của tập dữ liệu. File u.item lưu thông tin về phim.
- u.genre: File lưu danh sách 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 để lấy tập đặc trưng nước sản xuất, hãng phim, đạo diễn, diễn viên chính để làm tập đặc trưng phim.
- u.user: File lưu thông tin về những người dùng. Các hàng có cấu trúc chung : user id | age | gender | occupation | zip code. User id được sử dụng trong tập dữ liệu u.data.
- u.occupation: File lưu danh sách các nghề nghiệp. Đây là tập đặc trưng nội dung người dùng được dùng trong thử nghiệm phương pháp đề xuất.

### 4.2. 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 $O_i$[2, 3, 18].

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

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

Sai số dự đoán trên toàn tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số dự đoán cho mỗi khách hàng thuộc $U_{te}$.Giá trị MAE nhỏ thì phương pháp dự đoán có độ chính xác cao [2, 3, 18].

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

### 4.3. So sánh và đánh giá

Phương pháp học bán giám sát đề xuất trong Mục 3 được thử nghiệm và so sánh với những phương pháp sau:

- Phương pháp KNN dựa vào người dùng sử dụng độ tương quan Pearson (ký hiệu là CF-UserBased). Đây là phương pháp tư vấn cộng tác chuẩn dựa vào người dùng được đề xuất trong [9].
- Phương pháp KNN dựa vào sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CF-ItemBased). Đây là phương pháp tư vấn cộng tác chuẩn dựa vào sản phẩm được đề xuất trong [10].
- Phương pháp KNN dựa vào hồ sơ người dùng sử dụng độ tương quan Pearson (ký hiệu là CBF-UserBased). Đây là phương pháp tư vấn dựa vào việc so sánh mức độ tương tự giữa hai hồ sơ người dùng được đề xuất theo công thức (15).
- Phương pháp KNN dựa vào hồ sơ sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CBF-ItemBased). Đây là phương pháp tư vấn dựa vào việc so sánh mức độ tương tự giữa hai hồ sơ sản phẩm được đề xuất theo công thức (26).
- Phương pháp tư vấn kết hợp KNN dựa vàongười dùngvà 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, 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 rất thưa [3]. Tập test4.inp có số đánh giá giá biết trước là 20 tương ứng với trường hợp dữ liệu huấn luyện thưa [3]. Chọn θ = 4, 8, 12, 15 ứng với mỗi bộ test 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 α=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 8 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 [1, 2, 3].

**Bảng 8.** 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** |

Kết quả trong Bảng 8 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 MAElần 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, 2].

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-UserBasedvà CF-UserBased cho lại giá trị MAE lần lượt là 0.865, 0.859, 855, 0.835và 0.824, 0.817, 0.821, 0.813 so với 0.793, 0.792, 0.791, 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, 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 VÀ HƯỚNG PHÁT TRIỂN

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. Su X., Khoshgoftaar T. M., "*A Survey of Collaborative Filtering Techniques.*,". Advances in Artificial Intelligence ,2009, pp.1-20.
2. Robin D. Burke, "*Hybrid Recommender Systems: Survey and Experiments*". User Model. User-Adapt. Interact. 12(4): 331-370 (2002).
3. Asela Gunawardana, Guy Shani, "*A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*. Journal of Machine Learning Research 10: 2935-2962 (2009).
4. Asela Gunawardana, Christopher Meek, " *A unified approach to building hybrid recommender systems*". RecSys 2009: 117-124.
5. Robin D. Burke, Fatemeh Vahedian, Bamshad Mobasher, "*Hybrid Recommendation in Heterogeneous Networks*". UMAP 2014: 49-60.
6. Pazzani, M. J. "*A framework for collaborative, content-based and demographic filtering*", Artificial Intelligence Review 13(5-6), 393–408 (1999).
7. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M. "*Combining content-based and collaborative filters in an online newspaper*". In: Proceedings of ACM SIGIR workshop on recommender systems, vol. 60. Citeseer (1999).
8. M. D. Ekstrand, J. T. Riedl and J. A. Konstan, "*Collaborative Filtering Recommender System*". Foundations and Trends in Human–Computer Interaction, Vol 4, No2, 2010, pp 81:173.
9. Breese J. S., Heckerman D., and Kadie C., "*Empirical analysis of Predictive Algorithms for Collaborative Filtering*", In Proc. of 14th Conf. on Uncertainty in Artificial (1998).
10. Sarwar B., Karypis G., Konstan J., and Riedl J., "*Item-Based Collaborative Filtering Recommendation Algorithms*", Proc. 10th Int'l WWW Conf (2001).
11. Nguyen Duy Phuong, Le Quang Thang, Tu Minh Phuong, *"A Graph-Based Method for Combining Collaborative and Content-Based Filtering.* PRICAI 2008: 859-869.
12. Nguyen Duy Phuong, Tu Minh Phuong, "Collaborative Filtering by Multi-task Learning", RIVF 2008, pp: 227-232.
13. Do Thi Lien, Nguyen Duy Phuong, "*Collaborative Filtering with a Graph-based Similarity Measure".* ComManTel, 2014, pp. 251-256.
14. Do Thi Lien, Nguyen Xuan Anh, Nguyen Duy Phuong, "*A Graph Model for Hybrid Recommender Systems".* KSE 2015, pp. 138-143.
15. Tran Nhat Quang, Do Thi Lien, Nguyen Duy Phuong, " *Collaborative Filtering by Co-training Method".* KSE 2014, pp. 273-285.
16. J. Wang, A. P. de Vries, and M. J. T. 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 (SIGIR '06). ACM, New York, NY, USA, 501-508.
17. Raghavan, S., Gunasekar, S., Ghosh, J. "*Review quality aware collaborative filtering*". In Proceedings of the sixth ACM conference on Recommender systems, pp. 123–130. ACM(2012).
18. Herlocker J. L., Konstan J. A., Terveen L. G., and Riedl J. T., "*Evaluating Collaborative Filtering Recommender Systems*", ACM Trans. Information Systems, vol. 22, No. 1 (2004), pp. 5-53.
19. http://www.grouplens.org/


# ASEMI-SUPERVISED LEARNING METHOD FOR HYBRID FILTERING

**Do Thi Lien, Nguyen Duy Phương**

***ABSTRACT***— *Recommender systems are the auto systems of providing appropriate information and removing unappropriate information for users. The recommender systems are built based on two main information filtering techniques: Collaborative filtering and content-based filtering.Content-based filtering perform effectively with information in text form but had difficulty in features seletion with multimedia information. Collaborative filtering perform well on all types of information but had problems when sparse data, new uses and new items.In this paper, we propose a new unify model between collaborative filtering and content-based filtering by a semi-supervised learning method. The model is built based on two semi-supervised procedures: the first procedure semi-supervise ratings set between users and item's features, the second procedure semi-supervise ratings set between items and user's features. The first procedure allows usto detect new items that is high suitable capability with the users. The second procedure allows us to detect new users that is high suitable ability with the items. Two procedures performed simultaneously and complement each other for suitable predicted values to improve recommender results. The experimental results on real data sets show that the proposed methods utilize effectively the advantages and limit disadvantages significantly of baseline filtering methods.*

***Keywords***—*Collaborative filtering recommendation, content-based filtering recommendation, hybrid filtering recommendation system, supervised learning recommendation, unsupervised learning recommendation, semi-supervised learning recommendation.*