

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

Unifying Collaborative and Content-based Filtering by Semi-Supervised Learning Method

Đỗ Thị Liên, Nguyễn Duy Phương, Từ Minh Phương

Abstract. Recommender systems are the auto systems of providing appropriate information and removing inappropriate 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 selection with multimedia information. Collaborative filtering perform well on all types of information but had problems when sparse data, new users and new items. In this paper, we propose a new unify model between collaborative filtering and content-based filtering by a co-training method. The experimental results on real data sets show that the proposed methods utilize effectively the advantages and limit disadvantages significantly of baseline filtering methods.

Keyword: 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 các sản phẩm (trang web, bản tin, phim, video...) phù hợp cho mỗi người dùng. Hệ tư vấn được xây dựng dựa trên tập gồm N người dùng $U = \{u_1, u_2, \dots, u_N\}$, $P = \{p_1, p_2, \dots, p_M\}$ là tập gồm M sản phẩm. Mỗi sản phẩm $x \in P$ 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. Mỗi quan hệ giữa tập người dùng U và tập sản phẩm P được biểu diễn thông qua ma trận đánh giá $R = \{r_{ix}; i = 1, 2, \dots, N; x = 1, 2, \dots, M\}$. Giá trị r_{ix} thể hiện đánh giá của người dùng $u_i \in U$ cho một số sản phẩm $x \in P$. Thông thường giá trị r_{ix} nhận một giá trị thuộc miền $F = \{1, 2, \dots, g\}$ được thu thập

trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị $r_{ix} = 0$ được hiểu người dùng u_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 $r_{ix} \neq 0$ nhỏ hơn 1%, hầu hết các giá trị r_{ix} còn lại là 0 [2, 3]. 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 mới phù hợp cho người dùng i [1, 5].

Hệ tư vấn cộng tác được tiếp cận theo hai phương pháp chính: phương pháp tư vấn dựa vào người dùng (user-based) và tư vấn dựa vào sản phẩm (item-based). Dựa vào các độ đo tương tự, mỗi phương pháp được thực hiện thông qua bốn bước: xác định mức độ tương tự giữa các cặp người dùng hoặc sản phẩm, tìm tập láng giềng của người dùng hoặc sản phẩm, dự đoán mức độ phù hợp của người dùng hoặc sản phẩm dựa trên tập láng giềng, và sinh ra tư vấn [2, 12, 13]. Do tập dữ liệu đánh giá của người dùng đối với sản phẩm thưa nên cả hai phương pháp gặp phải một số hạn chế dưới đây:

- (i) Việc tính toán mức độ tương tự giữa các cặp người dùng dựa trên tập các đánh giá giao nhau nên hai người dùng có đánh giá giao nhau rất ít nhưng lại giống nhau về giá trị đánh giá thì hai người dùng này được xác định là hoàn toàn tương tự nhau. Nhiều cặp người dùng có đánh giá không giao nhau bất kỳ sản phẩm nào thì hai người dùng này được xác định là hoàn toàn không tương tự nhau. Cả hai điều này khiến cho việc xác định tập láng giềng, dự đoán và sinh ra tư vấn của các phương pháp user-based có kết quả thấp.
- (ii) Việc tính toán mức độ tương tự giữa các cặp sản phẩm cũng dựa trên tập các đánh giá giao nhau nên hai sản phẩm có đánh giá giao nhau rất ít nhưng lại giống nhau về giá trị đánh giá thì hai sản phẩm này được xác định là hoàn toàn tương tự nhau. Nhiều cặp sản phẩm có đánh giá không giao nhau bất kỳ người dùng nào thì hai sản phẩm này được xác định là hoàn toàn không tương tự nhau. Cả hai điều này khiến cho việc xác

định tập láng giềng, dự đoán và sinh ra tư vấn của các phương pháp item-based cũng có kết quả thấp.

Trong quá trình nghiên cứu thực nghiệm trên các tập dữ liệu thực, chúng tôi phát hiện thấy nếu chỉ quan sát theo người dùng thì nhiều cặp người dùng gặp phải vấn đề dữ liệu thừa, nhưng khi quan sát theo sản phẩm thì tập sản phẩm chưa được người dùng đánh giá không hẳn đã thừa. Trong trường hợp này dự đoán theo phương pháp item-based có thể phát huy tác dụng tốt hơn phương pháp user-based [7, 8, 9]. Ngược lại, khi quan sát theo sản phẩm, nhiều cặp sản phẩm gặp phải vấn đề dữ liệu thừa, nhưng khi quan sát theo người dùng thì tập người dùng chưa đánh giá sản phẩm không hẳn đã thừa. Trong trường hợp này dự đoán theo phương pháp user-based có thể phát huy tác dụng tốt hơn phương pháp item-based [7, 8, 9].

Một khía cạnh quan trọng khác được phát hiện trong khi thực hiện độc lập phương pháp user-based và item-based là dự đoán theo phương pháp item-based sẽ làm cho một số cặp người dùng không gặp phải vấn đề dữ liệu thừa. Ngược lại, dự đoán theo phương pháp user-based sẽ làm cho một số cặp sản phẩm không gặp phải vấn đề dữ liệu thừa. Chính vì vậy, việc hợp nhất giữa hai phương pháp user-based và item-based chắc chắn sẽ cải thiện được chất lượng dự đoán và hạn chế được vấn đề dữ liệu thừa của hệ tư vấn.

Trong bài báo này, chúng tôi đề xuất một mô hình hợp nhất giữa phương pháp user-based và item-based cho lọc kết hợp bằng kỹ thuật học bán giám sát nhằm tận dụng lợi thế và hạn chế vấn đề dữ liệu thừa của mỗi phương pháp. Thuật toán được xây dựng dựa trên hai thủ tục: 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 người dùng. 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.

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 cùng là kết luận và hướng phát triển tiếp theo của bài báo.

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

Không hạn chế tính tổng quát, 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 $C = \{c_1, c_2, \dots, c_{|C|}\}$ được xác định theo công thức (2). Mỗi người dùng $i \in U$ được biểu diễn thông qua $|T|$ đặc trưng nội dung $T = \{t_1, t_2, \dots, t_{|T|}\}$ được xác định theo công thức (3).

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

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

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

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 [7, 12]. Hệ tư vấn nội dung được xây dựng dựa trên ma trận các đặc trưng nội dung C và T [10, 11]. Hệ tư vấn lai xây dựng dựa trên ma cả ba ma trận R , C và T [3, 4, 5].

Bảng 1. 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 2. Ma trận đặc trưng sản phẩm C

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

Bảng 3. 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

II.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 ta 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 [3, 4, 6, 11]. Gọi $P_i \subseteq P$ được

xác định theo (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 $i \in U$.

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

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.

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

Để hạn chế một số đặc trưng nội dung ít được người dùng quan tâm nhưng vẫn được đánh giá với trọng số cao chúng tôi thực hiện quan sát trên tập $Item(i, s)$ nếu $|Item(i, s)|$ 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 .

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

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 $s \in C$. Dễ dàng nhận thấy $w_{is} \in F$, trong đó $F = \{1, 2, \dots, g\}$. 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 (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.

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

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 $\theta = 2$, 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 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 [3, 4, 5].

II.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 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 [3]. 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.

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

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 $|Item(i, s)|$ 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_{qx} được tính bằng trung bình cộng của tất cả các giá trị đánh giá. Trường hợp $|User(x, q)|$ có giá trị bé hơn θ , giá trị v_{qx} được tính bằng tổng của tất cả các giá trị đánh giá chia cho θ .

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

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. Vì vậy ta có thể xem mỗi đặc trưng nội dung người dùng đóng vai trò như một người dùng phụ

bổ sung vào tập người dùng. Dựa trên nhận xét này, 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 sản phẩm phụ bổ sung vào để mở rộng ma trận đánh giá về phía sản phẩm.

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

Ví dụ với hệ có ma trận đánh giá theo Bảng 1, ma trận đặc trưng người dùng theo Bảng 3, chọn $\theta = 2$, 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 kết hợp theo sản phẩm đều có thể dễ dàng triển khai trên ma trận đánh giá mở rộng theo hồ sơ sản phẩm [3, 13].

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, 3]. Để khắc phục điều này, chúng tôi đề xuất thuật toán tư vấn kết hợp bằng phương pháp học bán giám sát. Phương pháp được tiến hành như dưới đây.

III.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 (12) để giám sát việc tính toán mức độ tương tự giữa các cặp người dùng. Trong đó, P_i được xác định theo (4), C_i được xác định theo (13).

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

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

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 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 thừa. Dựa vào S_i và độ tương quan Pearson [4, 9], 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).

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

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

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

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

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

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

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

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

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 thừa. 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 thừa. 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 thừa 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

Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT

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)[1, 12].

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

$$r_{ix} = \bar{r}_i + \frac{\sum_{j \in K_i} (r_{jx} - \bar{r}_j) u_{ij}}{\sum_{j \in K_i} |u_{ij}|} \quad (22)$$

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.

III.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 (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).

$$S_x = \{y \in P: |U_x \cap U_y| \geq \gamma_1 \text{ và } |T_x \cap T_y| \geq \gamma_2\} \quad (23)$$

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

S_x được xác định theo (23) là tập sản phẩm $y \in P$ có số lượng người dùng đánh giá 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 thừa. 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).

$$a_{xy} = \begin{cases} 0 & \text{nếu } y \notin S_x \\ \frac{\sum_{i \in U_x \cap U_y} (r_{ix} - \bar{r}_x)(r_{iy} - \bar{r}_y)}{\sqrt{\sum_{i \in U_x \cap U_y} (r_{ix} - \bar{r}_x)^2} \sqrt{\sum_{i \in U_x \cap U_y} (r_{iy} - \bar{r}_y)^2}} & \text{nếu } y \in S_x \end{cases} \quad (25)$$

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

$$p_{xy} = \begin{cases} \frac{\sum_{i \in H_x \cap H_y} (r_{ix} - \bar{r}_x)(r_{iy} - \bar{r}_y)}{\sqrt{\sum_{i \in H_x \cap H_y} (r_{ix} - \bar{r}_x)^2} \sqrt{\sum_{i \in H_x \cap H_y} (r_{iy} - \bar{r}_y)^2}} & \text{nếu } y \in S_x \text{ và } a_{xy} \geq \alpha \text{ và } b_{xy} \geq \alpha \\ 0 & \text{trung các trường hợp khác} \end{cases} \quad (27)$$

Trong đó, U_x được xác định theo công thức (8), T_x được xác định theo công thức (24), H_x , $\bar{r}_x$, $\bar{r}_x^*$, $\bar{r}_x^*$ được xác định theo công thức (28), (29), (30), (31), theo thứ tự.

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

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

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

$$\bar{r}_x^* = \frac{1}{|H_x \cap H_y|} \sum_{i \in H_x \cap H_y} r_{ix} \quad (31)$$

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 thừa. 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 thừa. Giá trị p_{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 thừa 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 $\alpha=0.90$ 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)[1, 9, 13].

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

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

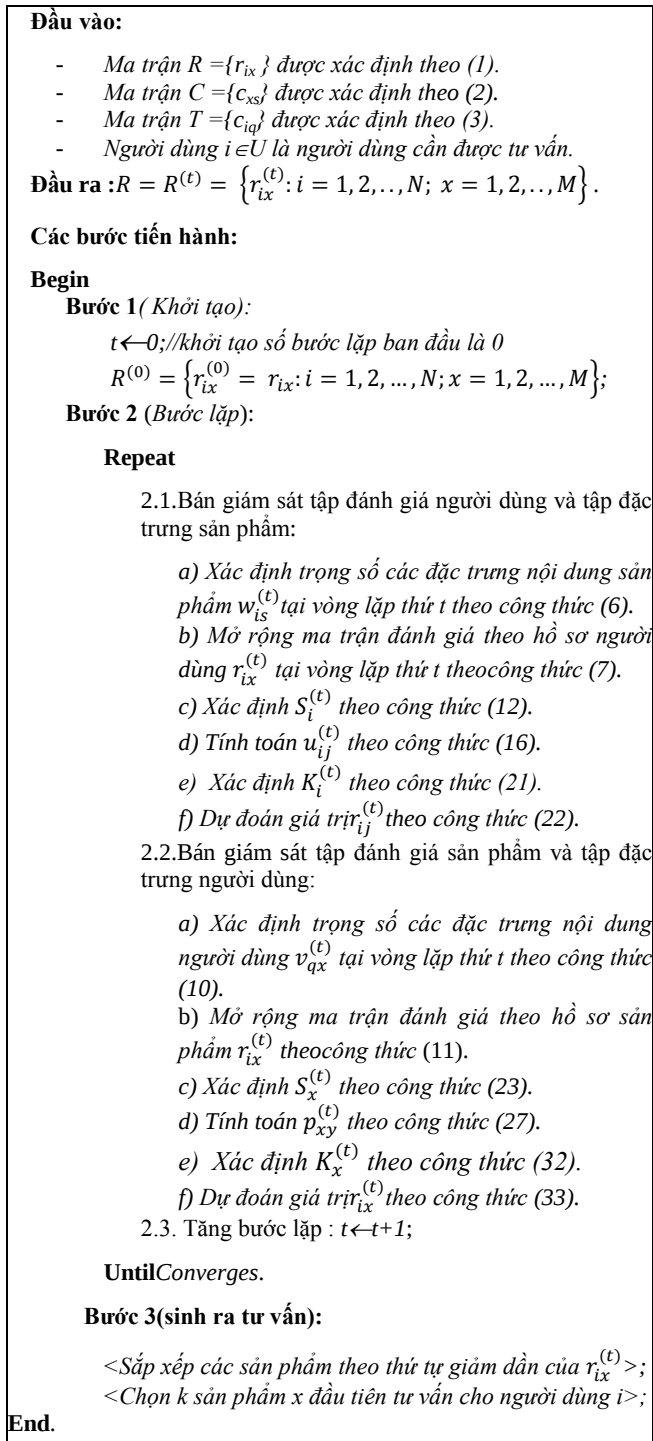
Giá trị dự đoán r_{ix} theo (33) phản ánh mức độ phù hợp của người dùng $i \in U$ đối với sản phẩm $x \in P$ được bổ sung vào ma trận đánh giá mở rộng theo sản phẩm để phục vụ quá trình bán giám sát theo tập đánh giá người dùng và tập đặc trưng sản phẩm. Hai quá trình bán giám sát được thực hiện đồng 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, 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.

III.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

Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT

trung người dùng. Thuật toán được mô tả chi tiết như trong Hình 1.



Hình 1. Thuật toán học bán giám sát

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 thừa 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 dùng 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 .

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.

IV.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 100KB, MovieLens 1MB và MovieLens 10MB. Trong đó, tập dữ liệu MovieLens 100KB là tập con của tập MovieLens 1MB. Tập dữ liệu MovieLens 1MB cung cấp đầy đủ tập đặc trưng sản phẩm và người dùng kèm theo tập đánh giá người dùng. Tập dữ liệu MovieLens 10M tuy lớn nhưng không cung cấp tập đặc trưng người dùng và tập đặc trưng sản phẩm. Chính vì vậy,

chúng tôi sử dụng tập dữ liệu MovieLens 1M 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 file sau [14]:

- 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. Chúng tôi tách trong IMDB để lấy tập đặc trưng nước sản xuất, hãng phim, đạo diễn để làm tập đặc trưng phim. Chúng tôi tách được 46 nước, mỗi nước có tối thiểu 20 phim; 72 hãng sản xuất, mỗi hãng giới thiệu ít nhất 20 phim; 734 đạo diễn cho tất các các 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.

IV.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 [4, 9].

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 .

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

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 [3, 9].

$$MAE = \frac{\sum_{u \in U_{te}} MAE_u}{|U_{te}|} \quad (35)$$

IV.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) [1, 12].
- Phương pháp tư vấn cộng tác dựa vào sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CF-ItemBased) [1, 13].
- Phương pháp tư vấn nội dung dựa vào hồ sơ người dùng sử dụng độ tương quan Pearson (ký hiệu là CBF-UserBased) [11].
- Phương pháp tư vấn nội dung dựa vào hồ sơ sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CBF-ItemBased) [10].
- Phương pháp tư vấn hợp nhất giữa user-based và item-based sử dụng độ tương quan Pearson (ký hiệu là Unify-UserItem) [8]. Đây là phương pháp tư vấn các sản phẩm có mức độ tương tự cao của cả hai phương pháp user-based và item-based.
- 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, test2.inp, test3.inp, test3.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 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 biết trước là 20 tương ứng với trường hợp dữ liệu huấn luyện tương đối đầy đủ. Chọn $\theta = 4, 8, 12, 15$ ứng với mỗi bộ test theo thứ tự để xác định xác định w_{is}, v_{qx} theo công thức (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 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 [3, 9, 7].

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
UNIFY-USERITEM	0.812	0.798	0.775	0.774
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 CBF-UserBased, CBF-ItemBased 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 CF-UserBased, CF-ItemBased cho lại giá trị MAE nhỏ hơn so với các phương pháp tư vấn theo nội dung. Phương pháp Unify-UserItem cho lại giá trị MAE tốt hơn hai phương pháp CF-UserBased, CF-ItemBased. Những kết quả này hoàn toàn phù hợp với những nghiên cứu trước đây [1, 2, 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, 855, 0.835 và 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 (*giá trị được in đậm trong Bảng 8*). 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 thừa đối với người dùng và sản phẩm là hoàn toàn tin cậy. Phương pháp chuyên 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 chuyên 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 thừa 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 thừa đối với mỗi sản phẩm. Dựa trên các tập không thừa 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 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 chuyên 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. 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.
3. ROBIN D. BURKE, "Hybrid Recommender Systems: Survey and Experiments", User Model. User-Adapt. Interact. 12(4): 331-370 (2002).
4. ASELA GUNAWARDANA, GUY SHANI, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks", Journal of Machine Learning Research 10: 2935-2962 (2009).
5. ASELA GUNAWARDANA, CHRISTOPHER MEEK, "A unified approach to building hybrid recommender systems". RecSys 2009: 117-124.
6. ROBIN D. BURKE, FATEMEH VAHEDIAN, BAMSHAD MOBASHER, "Hybrid Recommendation in Heterogeneous Networks", UMAP 2014: 49-60.
7. RAGHAVAN, S., GUNASEKAR, S., GHOSH, J., "Review quality aware collaborative filtering", Proceedings of the sixth ACM conference on Recommender systems, pp. 123-130. ACM(2012).
8. J. WANG, A. P. DE VRIES, AND M. J. T. REINDERS, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion.", 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.
9. 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.
10. PAZZANI, M.J., "A framework for collaborative, content-based and demographic filtering", Artificial Intelligence Review 13(5-6), 393-408 (1999).
11. CLAYPOOL, M., GOKHALE, A., MIRANDA, T., MURNIKOV, P., NETES, D., SARTIN, M., "Combining content-based and collaborative filters in an online newspaper", Proceedings of ACM SIGIR workshop on recommender systems, vol. 60. Citeseer (1999).
12. BREESE J. S., HECKERMAN D., AND KADIE C., "Empirical analysis of Predictive Algorithms for Collaborative Filtering", Proceedings of 14th Conf. on Uncertainty in Artificial (1998).
13. SARWAR B., KARYPIS G., KONSTAN J., AND RIEDL J., "Item-Based Collaborative Filtering Recommendation Algorithms", Proceedings 10th Int'l WWW Conf (2001).
14. <http://www.grouplens.org/>

SƠ LƯỢC VỀ TÁC GIẢ

ĐỖ THỊ LIÊN



Sinh ngày 29/11/1987 tại Hà Nội.

Tốt nghiệp đại học và thạc sỹ tại Học Viện Bưu Chính Viễn Thông vào các năm 2010 và 2013.

Hiện là giảng viên tại Học Viện Bưu Chính Viễn Thông.

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

Email: liendt@ptit.edu.vn

NGUYỄN DUY PHƯƠNG



Sinh ngày 20/02/1965 tại Hà Nội.

Tốt nghiệp đại học và thạc sỹ tại trường Đại học Tổng hợp Hà Nội vào các năm 1988 và 1997. Bảo vệ tiến sỹ tại đại học Quốc Gia Hà Nội năm 2010.

Hiện là Phó trưởng khoa Công nghệ thông tin, Học viện CN Bưu chính Viễn thông.

Lĩnh vực nghiên cứu: học máy ứng dụng trong lọc thông tin.

Email: phuongnd@ptit.edu.vn

TỪ MINH PHƯƠNG



Sinh ngày 13/01/1971 tại Hà Nội.

Tốt nghiệp đại học Bách khoa Taskent năm 1993, bảo vệ tiến sỹ tại Viện hàn lâm khoa học Uzbekistan, Taskent năm 1995.

Hiện là Phó Giáo sư, Trưởng khoa Công nghệ thông tin, Học viện CN Bưu chính Viễn thông.

Lĩnh vực nghiên cứu: trí tuệ nhân tạo, học máy, tin sinh học.

Email: phuongtm@ptit.edu.vn