MỘT MÔ HÌNH ĐỒ THỊ CHO HỆ TƯ VẤN LAI

[1]Đỗ Thị Liên, [1]Nguyễn Xuân Anh, [1]Nguyễn Duy Phương, [1]Từ Minh Phương

[1]Học viện Công nghệ Bưu chính Viễn thông

liendt@ptit.edu.vn, anhnx@ptit.edu.vn,phuongnd@ptit.edu.vn, phuongtm@ptit.edu.vn

Tóm tắt - *Hệ tư vấn (recommender systems) là hệ thống có khả năng cung cấp thông tin phù hợp và gỡ bỏ thông tin không phù hợp cho mỗi người dùng sử dụng các dịch vụ Internet. 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). Mỗi phương pháp khai thác những khía cạnh riêng liên quan đến đặc trưng nội dung hay thói quen sử dụng sản phẩm của người dùng trong quá khứ để dự đoán một danh sách ngắn gọn các sản phẩm phù hợp nhất đối với mỗi người dùng. Trong bài báo này, chúng tôi đề xuất một phương pháp hợp nhất giữa tư vấn cộng tác và tư vấn nội dung bằng mô hình đồ thị. Mô hình cho phép ta dịch chuyển bài toán tư vấn kết hợp về bài toán tư vấn cộng tác, sau đó xây dựng một độ tương tự mới trên đồ thị để xác định mức độ tương tự giữa các cặp người dùng và sản phẩm để sinh ra kết quả dự đoán. Kết quả thử nghiệm trên các bộ dữ liệu thực về phim cho thấy các phương pháp đề xuất cải thiện đáng kể chất lượng tư vấn.*

Từ khóa - Tư vấn cộng tác, tư vấn theo nội dung, hệ tư vấn lai, tư vấn dựa vào sản phẩm, tư vấn dựa vào người dùng.

I. GIỚI THIỆU HỆ TƯ VẤN

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 [4]. 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 ϕ [4]. Ma trận R chính là đầu vào của các hệ thống tư vấn cộng tác [1, 2, 3]. 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 $x \in P$. 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, 17]. 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 $x \in P$ 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 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, 13]. 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.

Có nhiều đề xuất khác nhau để giải quyết bài toán tư vấn. 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 $V=\{v_{iq}\}$ [3, 13, 17]. Hệ tư vấn cộng tác đưa ra phương pháp dự đoán dựa trên ma trận đánh giá $R=\{r_{ix}\}$ [1, 2, 4]. Hệ tư vấn lai đưa ra phương pháp dự đoán dựa trên cả 3 ma trận R, W và V[3, 9].

II. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

Hiệu quả của phương pháp tư vấn lai đã được khẳng định trong nhiều nghiên cứu khác nhau [2, 8]. Hướng tiếp cận phổ biến nhất thường được sử dụng là phương pháp tổ hợp tuyến tính giữa lọc cộng tác và lọc nội dung. Trong hướng tiếp cận này, các tác giả tiến hành xây dựng hai phương pháp lọc cộng tác và lọc nội dung độc lập nhau, sau đó tổ hợp tuyến tính kết quả dự đoán của cả hai hoặc lựa chọn ứng viên tốt nhất từ một trong hai phương pháp [17]. Hướng tiếp cận thứ hai xem xét vấn đề tư vấn lai bằng cách thêm các đặc trưng của lọc nội dung vào lọc cộng tác. Phương pháp được thực hiện bằng cách xây dựng một thủ tục kết hợp dữ liệu để tạo nên dữ liệu đầu vào tổng hợp giữa các giá trị đánh giá của lọc cộng tác và các đặc trưng nội dung. Pazzani [13] đề xuất phương pháp biểu diễn hồ sơ sản phẩm bằng một vector trọng số các đặc trưng nội dung người dùng. Dựa trên biểu diễn này, phương pháp dự đoán được Pazzani thực hiện bằng các kỹ thuật lọc cộng tác thuần túy. Hướng tiếp cận thứ ba xem xét bài toán tư vấn lai bằng cách thêm các đặc trưng của lọc cộng tác vào lọc nội dung. Theo phương pháp này, các đặc trưng nội dung sản phẩm đóng vai trò trung tâm và xem xét đánh giá người dùng của lọc cộng tác như các giá trị đặc trưng giả định để thêm vào quá trình dự đoán [17, 18].

Hướng tiếp cận cuối cùng được cộng đồng quan tâm nghiên cứu là hợp nhất giữa lọc cộng tác và lọc nội dung dựa trên các kỹ thuật học máy. Basu [19] đề xuất việc xây dựng tập các giá trị đặc trưng đại diện cho cả lọc cộng tác và lọc nội dung. Phương pháp dự đoán được tiến hành dựa trên việc xây dựng tập luật suy diễn các giá trị đặc trưng. Popescul [20] đề xuất mô hình phân tích ngữ nghĩa ẩn để hợp nhất giữa lọc cộng tác và lọc nội dung. Balisico và Hofman [21] sử dụng hàm nhân để kết hợp mức độ tương tự từ người dùng đến người dùng, sản phẩm đến sản phẩm, sau đó áp dụng máy vector hỗ trợ để sinh ra dự đoán. Crammer và Singer [22] xem xét bài toán tư vấn lai như việc xếp hạng các sản phẩm bằng việc bổ sung các đặc trưng nội dung sản phẩm.

Liên quan đến mô hình đồ thị, nhiều đề xuất khác nhau đã được đưa ra giải quyết bài toán tư vấn. Aggarwal [23] biểu diễn mối quan hệ giữa các cặp người dùng như một đồ thị có hướng, trong đó mỗi cạnh được thiết lập phản ánh mức độ tương tự giữa hai người dùng. Phương pháp dự đoán được thực hiện bằng cách tính toán trọng số đường đi ngắn nhất giữa các cặp người dùng. Lien [7] đề xuất xây dựng độ đo tương tự giữa các cặp người dùng hoặc sản phẩm bằng mô hình đồ thị hai phía có trọng số. Mức độ tương tự giữa các cặp người dùng được thực hiện bằng cách ước lượng tổng trọng số của tất cả các đường đi từ đỉnh người dùng đến đỉnh người dùng, mức độ tương tự giữa các cặp sản phẩm được thực hiện bằng cách ước lượng tổng trọng số của tất cả các đường đi từ sản phẩm dùng đến đỉnh sản phẩm. Phuong [6] đề xuất phương pháp kết hợp giữa lọc cộng tác và lọc nội dung bằng cách xây dựng mối liên hệ giữa người dùng và tập đặc trưng nội dung sản phẩm. Phương pháp dự đoán được thực hiện bằng cách tổ hợp tuyến tính trọng số các đường đi từ đỉnh người dùng đến đỉnh sản phẩm. Sản phẩm nào có tổng trọng số các đường đi lớn nhất đến nó chính là đích của quá trình dự đoán.

Trong bài báo này, chúng tôi đề xuất một mô hình hợp nhất giữa lọc cộng tác và lọc nội dung dựa trên biểu diễn đồ thị. Mô hình được xây dựng bằng cách lấy lọc cộng tác làm trung tâm, xây dựng hồ sơ người dùng dựa trên ma trận đánh giá để thiết lập nên mối quan hệ trực tiếp giữa tập người dùng với tập đặc trưng nội dung sản phẩm. Tiếp đến, chúng tôi tiến hành xây dựng hồ sơ sản phẩm cũng dựa trên ma trận đánh giá để thiết lập nên mối quan hệ trực tiếp giữa tập sản phẩm và tập đặc trưng nội dung người dùng. Dựa trên mối quan hệ giữa tập người dùng với tập đặc trưng nội dung sản phẩm và mối quan hệ giữa tập sản phẩm với tập đặc trưng nội dung người dùng, chúng tôi tìm cách xác định được mối quan hệ tiềm ẩn giữa tập đặc trưng sản phẩm và tập đặc trưng người dùng. Bằng cách này, chúng tôi thu gọn mô hình tư vấn kết hợp tổng quát thành mô hình tư vấn cộng tác chuẩn.

Về nguyên tắc, sau khi thu được mô hình tư vấn cộng tác chuẩn ta có thể triển khai bất kỳ một phương pháp lọc cộng tác nào đã được đề xuất trước đây. Tuy nhiên để khai thác được thế mạnh của đồ thị, chúng tôi xây dựng một độ tương tự dựa trên đồ thị bằng cách ước lượng mức độ tương tự giữa các cặp người dùng dựa trên tổng trọng số các đường đi từ đỉnh người dùng đến đỉnh người dùng, mức độ tương tự giữa các cặp sản phẩm dựa trên tổng trọng số các đường đi từ đỉnh sản phẩm đến đỉnh sản phẩm. Bằng cách này ta tận dụng được các thuật toán tìm kiếm hiệu quả đã được triển khai trên đồ thị. Để tập trung trình bày phương pháp đề xuất, mục tiếp theo chúng tôi trình bày bày phương pháp dịch chuyển bài toán tư vấn kết hợp về bài toán lọc cộng tác. Mục 4 trình bày về các phương pháp tư vấn lai dựa trên đồ thị. Mục 5 trình bày phương pháp thử nghiệm và so sánh. Mục cuối cùng là kết luận và hướng phát triển tiếp theo của bài báo.

III. DỊCH CHUYỂN BÀI TOÁN TƯ VẤN KẾT HỢP VỀ BÀI TOÁN LỌC CỘNG TÁC

Như đã giới thiệu ở trên, bài toán tư vấn kết hợp thực hiện dự đoán dựa trên tập đánh giá của người dùng cho các sản phẩm, cùng với tập đặc trưng nội dung sản phẩm và đặc trưng người dùng. Trong mục này, chúng tôi đề xuất phương pháp dịch chuyển bài toán tư vấn kết hợp về bài toán tư vấn cộng tác thuần túy bằng cách xây dựng hồ sơ người dùng và hồ sơ sản phẩm của dựa vào đánh giá tự nhiên của người dùng đối với các sản phẩm trong quá khứ. Trên cơ sở hồ sơ người dùng và hồ sơ sản phẩm đã được xây dựng, chúng tôi tìm cách xác định mối quan hệ tiềm ẩm giữa tập đặc trưng nội dung người dùng và tập đặc trưng nội dung sản phẩm để thu được mô hình giống với mô hình

bài toán tư vấn cộng tác. Để diễn giải tính đúng đắn của phương pháp đề xuất, chúng tôi sử dụng mô hình đồ thị thực hiện cho bài toán tư vấn kết hợp.

3.1. Biểu diễn đồ thị cho lọc kết hợp

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).

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

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

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

Hệ tư vấn với ma trận đánh giá $R = \{r_{ix}: i=1, 2, .., N; x=1, 2, .., M\}$, ma trận đặc trưng nội dung sản phẩm $C=\{c_{xs}: x=1, 2, .., M; s=1, 2, .., |C|\}$, ma trận đặc trưng nội dung người dùng $T = \{t_{iq}: i=1, 2, .., N; q=1, 2, .., |T|\}$ có thể biểu diễn dưới dạng đồ thị trọng số $G = (\Omega, E)$, trong đó Ω là tập đỉnh và E là tập cạnh. Tập đỉnh Ω của đồ thị được xác định theo công thức (4) chính là hợp của tập người dùng U, tập sản phẩm P, tập đặc trưng nội dung sản phẩm C và tập đặc trưng nội dung người dùng T. Tập cạnh E của đồ thị bao gồm 3 loại cạnh: cạnh (i, x) nối giữa đỉnh người dùng với đỉnh sản phẩm, cạnh (x, s) nối đỉnh sản phẩm với đặc trưng nội dung sản phẩm, cạnh (i, q) nối giữa đỉnh người dùng với đỉnh đặc trưng nội dung của người dùng.

$$E = \begin{cases} e = (i, x) & \text{nếu } r_{ix} \neq 0 : i \ U, x \ P. \\ e = (x, s) & \text{nếu } c_{xs} \neq 0 : x \ P, s \ C. \\ e = (i, q) & \text{nếu } t_{iq} \neq 0 : i \ U, q \ T. \end{cases} \quad (5)$$

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

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. Khi đó, đồ thị biểu diễn cho bài toán tư vấn tổng quát được biểu diễn như Hình 1. Đồ thị được biểu diễn thành ba đồ thị con hai phía (bipartie praph). Đồ thị con hai phía ở giữa biểu diễn quan điểm của người dùng đối với các sản phẩm thông qua ma trận đánh giá $R = (r_{ix})$. Các cạnh nối giữa đỉnh người dùng $i \in U$ với đỉnh sản phẩm $x \in P$ được đánh trọng số là r_{ix}. Đồ thị con hai phía trên cùng biểu diễn mối quan hệ giữa sản phẩm với tập đặc trưng nội dung sản phẩm thông qua ma trận $C = (c_{xs})$. Các cạnh nối giữa đỉnh sản phẩm $x \in P$ với đỉnh đặc trưng nội dung sản phẩm $s \in C$ được đánh trọng số là *1*. Đồ thị con hai phía dưới cùng biểu diễn mối quan hệ giữa người dùng với tập đặc trưng nội dung người dùng thông qua ma trận $T = (t_{iq})$. Các cạnh nối giữa đỉnh người dùng $i \in U$ với đỉnh đặc trưng nội dung người dùng $q \in T$ cũng được đánh trọng số là *1*.

Dựa trên biểu diễn đồ thị, phương pháp tư vấn cộng tác được thực hiện dựa trên các cạnh nối giữa đỉnh người dùng $i \in U$ và đỉnh sản phẩm $x \in P$ với trọng số r_{ix} [5]. Phương pháp tư vấn theo nội dung sản phẩm được thực hiện trên các cạnh nối giữa đỉnh sản phẩm $x \in P$ và đỉnh đặc trưng nội dung sản phẩm $s \in C$ [7]. Phương pháp tư vấn theo nội dung người dùng được thực hiện trên các cạnh nối giữa đỉnh người dùng $i \in U$ và đỉnh đặc trưng nội dung người dùng $t \in T$ [17]. Phương pháp tư vấn kết hợp được thực hiện dựa trên cả ba loại cạnh (i, x), (x, s), và (i, q) [9, 10].

3.2. Xây dựng hồ sơ người dùng dựa trên 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 nội dung không đóng góp vào việc xác định mức độ tương tự giữa vector hồ sơ người dùng và vector đặc trưng sản phẩm vẫn được tham gia quá trình tính toán [3, 5]. Để hạn chế điều này, chúng tôi đề xuất phương pháp xây dựng hồ sơ sử dụng sản phẩm của người dùng thông qua các giá trị đánh giá của hệ tư vấn cộng tác, sau đó thiết lập mối quan hệ trực tiếp giữa người dùng và từng đặc trưng sản phẩm để nâng cao hiệu quả tư vấn. Phương pháp được tiến hành như sau.

Để xây dựng được hồ sơ sử dụng 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, 17]. Gọi $P_i \subseteq P$ được xác định theo công thức (6) 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 $i \in U$.

$$P_i = \{x \in P \mid r_{ix} \neq 0 \ \ (i \in U, x \in P)\} \tag{6}$$

Gọi $ListItem(i, s)$ là tập các sản phẩm $x \in P_i$ chứa đựng đặc trưng $s \in C$ được xác định theo công thức (7). Khi đó, $|ListItem(i, s)|$ 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.

$$ListItem(i, s) = \{x \in P_i \mid c_{xs} \neq 0 \ \ (i \in U, x \in P, s \in C\} \tag{7}$$

Dựa trên P_i và $ListItem(i, s)$ 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 [17]. 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 [18], 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 phương pháp 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 $ListItem(i, s)$. Nếu giá trị $|ListItem(i, s)|$ 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 $|ListItem(i, s)|$ 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.

$$w_{is} = \begin{cases} \frac{1}{|ListItem(i,s)|} \sum_{x \in ListItem(i,s)} r_{ix} & nếu \ \ |ListItem(i,x)| \geq \theta \\ \frac{1}{\theta} \sum_{x \in ListItem(i,s)} r_{ix} & nếu \ \ |ListItem(i,x)| < \theta \end{cases} \tag{8}$$

Giá trị w_{is} được ước lượng theo (8) 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 mở rộng đồ thị hai phía của bài toán tư vấn cộng tác nguyên thủy (*đồ thị con ở giữa*) bằng cách giữ nguyên tập đỉnh ở phía người dùng U, tập đỉnh phía sản phẩm được mở rộng là $P \cup C$. Liên kết giữa đỉnh người dùng $i \in U$ và đỉnh sản phẩm $x \in P$ được thiết lập nếu $r_{ix} \neq 0$. Liên kết giữa đỉnh người dùng $i \in U$ và đỉnh đặc trưng sản phẩm $s \in C$ được thiết lập nếu $w_{is} \neq 0$. Ma trận đánh giá mở rộng được xác định theo công thức (9).

$$r_{ix} = \begin{cases} r_{ix} & nếu \ x \in P \ và \ r_{ix} \neq 0 \\ w_{is} & nếu \ s \in C \ và \ w_{is} \neq 0 \ (x = s) \end{cases} \tag{9}$$

Ví dụ với đồ thị biểu diễn hệ tư vấn kết hợp được cho trong Hình 1, chọn θ = 2 ta sẽ tính toán được ma trận đánh giá mở rộng trong Bảng 4 và đồ thị tư vấn cộng tác mở rộng được thể hiện như Hình 2. Các cạnh màu đỏ là những cạnh mới được bổ sung vào đồ thị hai phía của lọc cộng tác.

Bảng 4. Ma trận đánh giá mở rộng R

	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

3.3. Xây dựng hồ sơ sản phẩm dựa trên ma trận đánh giá

Tương tự như hồ sơ người dùng, hồ sơ sản phẩm lưu trữ lại vết tích 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 [3, 13]. Gọi $U_x \subseteq U$ được xác đinh theo công thức (10) 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 $x \in P$.

$$U_x = \{i \in U \mid r_{ix} \neq 0 \ \ (i \in U, x \in P)\} \tag{10}$$

Gọi *ListUser* (x, q) là tập người dùng $i \in U_x$ có đặc trưng $q \in T$ được xác định theo công thức (11). Khi đó, $|ListUser(x, q)|$ 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.

$$ListUser(x, q) = \{i \in U_x \mid t_{iq} \neq 0 \ \ (i \in U, x \in P, q \in T\} \tag{11}$$

Dựa trên U_x và *ListUser*(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 *ListUser*(x, q). Nếu giá trị $|ListItem(i, s)|$ 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_{xq} được tính bằng trung bình cộng của tất cả các giá trị đánh giá. Trường hợp $|ListUser(x, q)|$ 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 người dùng ít quan tâm đến sản phẩm nhưng vẫn được đánh giá với trọng số cao.

$$v_{xq} = \begin{cases} \frac{1}{|ListIUser(x,q)|} \sum_{i \in ListUser(x,q)} r_{ix} & nếu \ \ |ListUser(x, q)| \geq \theta \\ \frac{1}{\theta} \sum_{i \in ListUser(x,q)} r_{ix} & nếu \ \ |ListUser(x, q)| < \theta \end{cases} \tag{12}$$

Giá trị v_{xq} được ước lượng theo (12) 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 mở rộng đồ thị hai phía của bài toán tư vấn cộng tác đã được mở rộng trong Mục 2.2 bằng cách giữ nguyên tập đỉnh ở phía sản phẩm là $P \cup C$ và mở rộng phía người dùng thành $U \cup T$. Liên kết giữa đỉnh sản phẩm $x \in P$ và đỉnh người dùng $i \in U$ được thiết lập nếu $r_{ix} \neq 0$. Liên kết giữa đỉnh sản phẩm $x \in P$ và đỉnh đặc trưng người dùng $q \in T$ được thiết lập nếu $v_{xq} \neq 0$. Ma trận đánh giá mở rộng ghi lại trọng số các cạnh (x, i) và (x, q) được xác định theo công thức (13).

$$r_{ix} = \begin{cases} r_{ix} & nếu \ i \in U, x \in P \ và \ r_{ix} \neq 0 \\ w_{is} & nếu \ i \in U, s \in C \ và \ w_{is} \neq 0 \ (x = s) \\ v_{xq} & nếu \ x \in P, q \in T \ và \ v_{xq} \neq 0 \ (x = q) \end{cases} \tag{13}$$

Ví dụ với đồ thị biểu diễn hệ tư vấn kết hợp được cho trong Hình 1, chọn θ = 2 ta sẽ tính toán được ma trận đánh giá mở rộng trong Bảng 5 và đồ thị tư vấn cộng tác mở rộng được thể hiện như Hình 3. Các cạnh màu xanh là những cạnh mới được bổ sung vào đồ thị hai phía của lọc cộng tác.

Bảng 5. Ma trận đánh giá mở rộng R

	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
t_1	2	2	2	1			
t_2	0	0	2	0			
t_3	0	2	0	1			
t_4	2	2	4	0			

3.4. Xây dựng mối liên hệ giữa đặc trưng người dùng và đặc trưng sản phẩm

Hồ sơ người dùng được xác định theo (8), hồ sơ sản phẩm được xác định theo (12) được thực hiện dựa trên đánh giá tự nhiên của người dùng đối với sản phẩm và thói quen sử dụng sản phẩm của người dùng. Rõ ràng, bản thân tập đặc trưng nội dung người dùng và tập đặc trưng nội dung sản phẩm cũng tồn tại một mối quan hệ tự nhiên nào đó giữa hồ sơ người dùng và hồ sơ sản phẩm. Ví dụ tại sao trẻ em thích xem phim hoạt hình, nữ tuổi teen thích xem phim tình cảm, nam tuổi teen thích xem phim hành động…? Chúng tôi cho rằng khai thác được mối quan hệ tiềm ẩn kể trên sẽ cải thiện đáng kể chất lượng dự đoán các sản phẩm phù hợp với mỗi người dùng.

Để xác định mối liên hệ tiềm ẩn giữa đặc trưng $q \in T$ với đặc trưng $s \in C$, chúng tôi xây dựng hai kiểu quan sát khác nhau. Kiểu quan sát thứ nhất được tiến hành từ hồ sơ người dùng đến các đặc trưng nội dung sản phẩm. Kiểu quan sát thứ hai được thực hiện ngược lại từ hồ sơ sản phẩm đến các đặc trưng người dùng. Vì cả hai kiểu quan sát chỉ nhằm mục đính xác định mối quan hệ tiềm ẩn giữa cặp đặc trưng $q \in T$ với đặc trưng $s \in C$ nên chúng tôi tổ hợp kết quả giữa hai kiểu quán sát để thu được kết quả cuối cùng. Phương pháp cụ thể được tiến hành như sau.

Quan sát từ hồ sơ người dùng đến các đặc trưng nội dung sản phẩm

Gọi U_q là tập người dùng $i \in U$ có đặc trưng nội dung $q \in T$ được xác định theo công thức (14). Gọi $UserAttr(i,s)$ là tập người dùng $i \in U$ có đặc trưng $q \in T$ đã đánh giá các sản phẩm $x \in P$ chứa đựng đặc trưng $s \in C$ được xác định theo công thức (15). Khi đó, mối liên hệ giữa đặc trưng $q \in T$ và đặc trưng $s \in C$ được ước lượng theo công thức (16). Trong đó, w_{is} là hồ sơ người dùng $i \in U$ được xác định theo (8).

$$U_q = \{i \in U \mid t_{iq} \neq 0\} \tag{14}$$

$$UserAttr(q,s) = \{i \in U_q \mid w_{is} \neq 0\} \tag{15}$$

$$a_{qs} = \begin{cases} \frac{1}{|UserAttr(q,s)|} \sum_{i \in UserAttr(q,s)} w_{is} & nếu\ |UserAttr(q,s)| \geq \theta \\ \frac{1}{\theta} \sum_{i \in UserAttr(q,s)} w_{is} & nếu\ |UserAttr(q,s) < \theta \end{cases} \tag{16}$$

Giá trị a_{qs} được ước lượng theo (16) phản ánh mức độ ảnh hưởng của đặc trưng $s \in C$ lên tập người dùng có đặc trưng $q \in T$. Nếu số lượng tập người dùng $i \in U$ có đặc trưng $q \in T$ đã đánh giá các sản phẩm $x \in P$ chứa đựng đặc trưng $s \in C$ vượt quá ngưỡng θ thì a_{qs} được tính bằng giá trị trung bình của trọng số các đặc trưng s trong hồ sơ người dùng. Trong trường hợp khác, a_{qs} được tính tổng trọng số các đặc trưng s trong hồ sơ người dùng nhân với 1/θ. Bằng cách này chúng ta có thể hạn chế được các đặc trưng của người dùng hoặc sản phẩm ít được người dùng sử dụng nhưng được đánh giá với trọng số cao.

Quan sát từ hồ sơ sản phẩm đến các đặc trưng người dùng

Gọi P_s là tập sản phẩm $x \in P$ có đặc trưng nội dung $s \in C$ được xác định theo công thức (17). Gọi $ItemAttr(q, s)$ là tập sản phẩm có đặc trưng $s \in C$ đã được đánh giá bởi tập người dùng $i \in U$ chứa đựng đặc trưng $q \in T$ được xác định theo công thức (18). Khi đó, mức độ phù hợp của tập sản phẩm có đặc trưng s đối với tập người dùng $i \in U$ chứa đựng đặc trưng q theo công thức (18). Trong đó, v_{xq} là hồ sơ sản phẩm $x \in P$ được xác định theo (12).

$$P_s = \{x \in P \mid c_{xs} \neq 0\} \tag{17}$$

$$ItemAttr(q,s) = \{x \in P_s \mid v_{xq} \neq 0\} \tag{18}$$

$$b_{qs} = \begin{cases} \frac{1}{|ItemAttr(q,s)|}\sum_{x\in ItemAttr(q,s)} v_{xq} & nếu\ |ItemAttr(q,s)| \geq \theta \\ \frac{1}{\theta}\sum_{x\in ItemAttr(q,s)} v_{xq} & nếu\ |ItemAttr(q,s) < \theta \end{cases} \quad (19)$$

Giá trị b_{qs} được ước lượng theo (19) phản ánh mức độ ảnh hưởng của đặc trưng $q \in T$ lên tập sản phẩm có đặc trưng $s \in C$. Nếu số lượng sản phẩm $x \in P$ có đặc trưng $s \in C$ đã đánh giá các sản phẩm $i \in U$ chứa đựng đặc trưng $q \in T$ vượt quá ngưỡng θ thì b_{qs} được tính bằng giá trị trung bình của trọng số các đặc trưng q trong hồ sơ sản phẩm. Trong trường hợp khác, b_{qs} được tính tổng trọng số các đặc trưng q trong hồ sơ người dùng nhân với 1/θ. Bằng cách này chúng ta có thể hạn chế được các đặc trưng của người dùng hoặc sản phẩm ít được người dùng sử dụng nhưng được đánh giá với trọng số cao.

Tổng hợp giữa các kiểu quan sát

Như đã trình bày ở trên, giá trị a_{qs} được xác định theo (16) và b_{qs} được xác định theo (19) đều phản ánh thói quen tự nhiên sử dụng sản phẩm của tập người dùng có đặc trưng q đối với tập sản phẩm có đặc trưng s. Điểm khác biệt duy nhất giữa a_{qs} và b_{qs} là kiểu quan sát dựa vào hồ sơ người dùng hay hồ sơ sản phẩm. Để dung hòa giữa hai kiểu quan sát, chúng tôi chọn giá trị trung bình giữa a_{qs} và b_{qs} theo công thức (20). Trong đó, giá trị d_{qs} chỉ được thiết lập khi và chỉ khi a_{qs} và b_{qs} có giá trị khác 0. Điều này có nghĩa, mối quan hệ giữa đặc trưng $s \in C$ và đặc trưng $q \in T$ được thiết lập khi và chỉ khi các sản phẩm có đặc trưng s thực sự được nhiều người dùng quan tâm và ngược lại nhiều người dùng có đặc trưng q thực sự quan tâm đến các sản phẩm có đặc trưng s. Điều này là hoàn toàn phù hợp với tâm lý chung của người sử dụng sản phẩm.

$$d_{qs} = \begin{cases} \frac{1}{2}(a_{qs} + b_{qs}) & nếu\ a_{qs} \neq 0\ và\ b_{qs} \neq 0 \\ 0 & trong\ các\ trường\ hợp\ khác \end{cases} \quad (20)$$

Sau khi xác định được mối liên hệ giữa tập đặc trưng người dùng và tập đặc trưng sản phẩm, chúng tôi mở rộng đồ thị hai phía của bài toán tư vấn cộng tác đã được trình bày trong Mục 2.3 bằng cách bổ sung các liên kết giữa mỗi đặc trưng $s \in C$ với đặc trưng $q \in T$. Đồ thị cuối cùng ta nhận được có tập đỉnh là tập người dùng U, tập sản phẩm P, tập đặc trưng người dùng T và tập đặc trưng sản phẩm P. Tập đỉnh của đồ thị được chia thành hai phía, một phía là $U \cup T$, phía còn lại là $P \cup C$. Tập cạnh của đồ thị bao gồm 4 loại cạnh: cạnh (i, x) nối giữa đỉnh người dùng và đỉnh sản phẩm được đánh trọng số là r_{ix}, cạnh (i, s) nối giữa đỉnh người dùng và đỉnh đặc trưng sản phẩm được đánh trọng số là w_{is}, cạnh (q, x) nối giữa đỉnh đặc trưng người dùng và đỉnh sản phẩm được đánh trọng số là v_{qx}, cạnh (q, s) nối giữa đỉnh đặc trưng người dùng và đỉnh đặc trưng sản phẩm được đánh trọng số là d_{qs}.

$$r_{ix} = \begin{cases} r_{ix} & nếu\ r_{ix} \neq 0\ (i \in U\ và\ x \in P) \\ w_{is} & nếu\ w_{is} \neq 0\ (i \in U\ và\ x = s \in C) \\ v_{qx} & nếu\ v_{qx} \neq 0\ (i = q \in T\ và\ x \in P) \\ d_{qs} & nếu\ d_{qs} \neq 0\ (i = q \in T\ và\ x = s \in C) \end{cases} \quad (21)$$

Ví dụ với đồ thị biểu diễn hệ tư vấn kết hợp được cho trong Hình 1, chọn θ = 2 ta sẽ tính toán được ma trận đánh giá mở rộng trong Bảng 6 và đồ thị tư vấn cộng tác mở rộng được thể hiện như Hình 4. Các cạnh màu vàng là những cạnh mới được bổ sung vào đồ thị hai phía của lọc cộng tác.

Bảng 6. Ma trận đánh giá mở rộng R

	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
t_1	2	2	2	1	2	1	1
t_2	0	0	2	0	1	1	1
t_3	0	2	0	1	1	1	0
t_4	2	2	4	0	4	1	3

lai

Ma trận đánh giá mở rộng được đề xuất theo (21) đã tích hợp đầy đủ các giá trị đánh giá của lọc cộng tác, hồ sơ người dùng, hồ sơ sản phẩm, mối liên hệ giữa hồ sơ người dùng và hồ sơ nội dung sản phẩm của lọc nội dung. Trọng số các đặc trưng nội dung trong hồ sơ người dùng, hồ sơ sản phẩm và mối liên hệ giữa các đặc trưng nội dung có cùng metric với giá trị đánh giá. Chính vì vậy, các phương pháp tư vấn cộng tác dựa vào bộ nhớ [15, 16] hoặc các phương pháp tư vấn cộng tác dựa trên mô hình [6, 11, 12] đều có thể triển khai trên ma trận đánh giá mở rộng. Đây là đóng góp chính của bài báo trong xây dựng mô hình hợp nhất giữa tư vấn cộng tác và tư vấn theo nội dung.

IV. CÁC PHƯƠNG PHÁP DỰ ĐOÁN TRÊN ĐỒ THỊ KẾT HỢP

Sau khi dịch chuyển bài toán tư vấn kết hợp về bài toán lọc cộng tác chuẩn, về nguyên tắc ta có thể triển khai bất kỳ một phương pháp tư vấn cộng tác nào trên ma trận đánh giá mở rộng. Trong khuôn khổ của bài báo này, chúng tôi đề xuất mở rộng các phương pháp tư vấn cộng tác dựa vào bộ nhớ bằng cách mở rộng các độ tương quan trên trên ma trận đánh giá mở rộng. Sau đó, chúng tôi xây dựng một độ đo tương tự mới dựa vào các kỹ thuật tìm kiếm trên đồ thị. Kết quả thử nghiệm trên các bộ dữ liệu thực về phim cho thấy các phương pháp đề xuất cải thiện đáng kể kết quả tư vấn.

4.1. Phương pháp tư vấn kết hợp dựa vào người dùng

Phương pháp tư vấn cộng tác dựa vào người dùng (UserBased) thực hiện ước lượng mức độ tương tự giữa các cặp người dùng dựa vào các độ đo tương tự để từ đó sinh ra dự đoán các sản phẩm mới phù hợp với người dùng cần được tư vấn [12, 15]. Hiệu quả của phương pháp UserBased phụ thuộc vào tập giá trị đánh giá $R = (r_{ix})$ được xác định theo (1). Do tính chất thưa thớt của ma trận đánh giá nên việc xác định mức độ tương tự giữa các cặp người dùng gặp nhiều hạn chế [14]. Để khắc phục nhược điểm này, chúng tôi tiến hành mở rộng phương pháp tư vấn kết hợp trên ma trận đánh giá mở rộng R được xác định theo (21). Trong đó, việc ước lượng mức độ tương tự giữa các cặp người dùng không chỉ thực hiện trên ma trận đánh giá mà được mở rộng cho toàn bộ hồ sơ người dùng. Phương pháp được ký hiệu là Hybrid-UserBased và tiến hành thông qua bốn bước như dưới đây.

Bước 1. *Tính toán mức độ tương tự giữa các cặp người dùng*. Tại bước này ta có thể sử dụng các độ đo tương quan hoặc các độ đo tương tự để tính toán mức độ giống nhau giữa các cặp người dùng [15]. Gọi u_{ij} là mức độ tương tự giữa người dùng $i \in U$ và người dùng $j \in U$. Khi đó, độ tương quan Pearson giữa người dùng $i \in U$ và người dùng $j \in U$ được mở rộng trên tập đánh giá người dùng và hồ sơ người dùng theo công thức (22).

$$u_{ij} = \frac{\sum_{x \in P_i \cap P_j} (r_{ix} - \overline{r_i})\,(r_{jx} - \overline{r_j})}{\sqrt{\sum_{x \in P_i \cap P_j} (r_{ix} - \overline{r_i})^2} \sqrt{\sum_{x \in P_i \cap P_j} (r_{jx} - \overline{r_j})^2}} + \frac{\sum_{s \in C_i \cap C_j} (r_{is} - \ddot{r}_i)(r_{js} - \ddot{r}_j)}{\sqrt{\sum_{s \in C_i \cap C_j} (r_{is} - \ddot{r}_i)^2} \sqrt{\sum_{s \in C_i \cap C_j} (r_{js} - \ddot{r}_j)^2}} \tag{22}$$

Trong đó,

$$\overline{r_i} = \frac{1}{|P_i \cap P_j|} \sum_{x \in P_i \cap P_j} r_{ix} \tag{23}$$

$$\overline{r_j} = \frac{1}{|P_i \cap P_j|} \sum_{x \in P_i \cap P_j} r_{jx} \tag{24}$$

$$C_i = \{\, s \in C | r_{is} \neq 0 \,\} \tag{25}$$

$$\ddot{r}_i = \frac{1}{|C_i \cap C_j|} \sum_{s \in C_i \cap C_j} r_{is} \tag{26}$$

$$\ddot{r}_j = \frac{1}{|C_i \cap C_j|} \sum_{s \in C_i \cap C_j} r_{js} \tag{27}$$

Bước 2. *Xác định tập láng giềng cho người dùng cần tư vấn*. Tại bước này ta chỉ cần sắp xếp các giá trị u_{ij} theo thứ tự giảm dần, trong đó $i \in U$ là người dùng cần được tư vấn các sản phẩm $x \in P$. Sau đó chọn tập K người dùng đầu tiên làm tập láng giềng của người dùng i [15]. Ký hiệu tập láng giềng của người dùng $i \in U$ là K_i.

Bước 3. *Dự đoán quan điểm của người dùng đối với các sản phẩm mới*. Phương pháp phổ biến nhất để sinh ra dự đoán quan điểm của người dùng $i \in U$ cho sản phẩm mới $x \in P$ theo công thức (28)[15].

$$r_{ix} = \overline{r_i} + \frac{\sum_{j \in K_i} (r_{jx} - \overline{r_j}) u_{ij}}{\sum_{j \in K_i} |u_{ij}|} \tag{28}$$

Bước 4. *Chọn K sản phẩm mới có r_{ix} cao nhất tư vấn cho người dùng i.*

4.2. Phương pháp tư vấn kết hợp dựa vào sản phẩm

Phương pháp tư vấn cộng tác dựa vào sản phẩm (ItemBased) thực hiện ước lượng mức độ tương tự giữa các cặp sản phẩm dựa vào các độ đo tương tự để từ đó sinh ra dự đoán các sản phẩm mới phù hợp với người dùng cần được tư vấn [1, 2, 16]. Hiệu quả của phương pháp ItemBased phụ thuộc vào tập giá trị đánh giá người dùng $R = (r_{ix})$ được xác định theo (1). Do tính chất thưa thớt của ma trận đánh giá nên việc xác định mức độ tương tự giữa các cặp sản phẩm gặp nhiều hạn chế. Để khắc phục nhược điểm này, chúng tôi tiến hành mở rộng phương pháp tư vấn kết hợp trên ma trận đánh giá mở rộng R được xác định theo (21). Trong đó, việc ước lượng mức độ tương tự giữa các cặp sản phẩm không chỉ thực trên ma trận đánh giá mà được mở rộng cho toàn bộ hồ sơ sản phẩm. Phương pháp được viết tắt là Hybrid-ItemBased và tiến hành thông qua bốn bước như dưới đây.

Bước 1. *Tính toán mức độ tương tự giữa các cặp sản phẩm*. Tại bước này ta có thể sử dụng các độ đo tương quan hoặc các độ đo tương tự để tính toán mức độ giống nhau giữa các cặp sản phẩm [16]. Gọi p_{xy} là mức độ tương tự giữa sản phẩm $x \in P$ và sản phẩm $y \in P$. Khi đó, độ tương quan Pearson giữa sản phẩm $x \in P$ và sản phẩm $y \in P$ được mở rộng trên tập đánh giá người dùng và hồ sơ sản phẩm theo công thức (29).

$$p_{xy} = \frac{\sum_{i \in U_x \cap U_y} (r_{ix} - \overline{r_x})\,(r_{iy} - \overline{r_y})}{\sqrt{\sum_{i \in U_x \cap U_y} (r_{ix} - \overline{r_x})^2} \sqrt{\sum_{i \in U_x \cap U_y} (r_{iy} - \overline{r_y})^2}} + \frac{\sum_{q \in T_x \cap T_y} (r_{qx} - \ddot{r}_x)(r_{qy} - \ddot{r}_y)}{\sqrt{\sum_{q \in T_x \cap T_y} (r_{qx} - \ddot{r}_x)^2} \sqrt{\sum_{q \in T_x \cap T_y} (r_{qy} - \ddot{r}_y)^2}} \tag{29}$$

Trong đó,

$$\overline{r_x} = \frac{1}{|U_x \cap U_y|} \sum_{i \in U_x \cap U_y} r_{ix} \quad (30)$$

$$\overline{r_y} = \frac{1}{|U_x \cap U_y|} \sum_{i \in U_x \cap U_y} r_{iy} \quad (31)$$

$$T_x = \{ q \in T | r_{qx} \neq 0 \} \quad (32)$$

$$\ddot{r_x} = \frac{1}{|T_x \cap T_y|} \sum_{q \in T_x \cap T_y} r_{qx} \quad (33)$$

$$\dddot{r_y} = \frac{1}{|T_x \cap T_y|} \sum_{q \in T_x \cap T_y} r_{qy} \quad (34)$$

Bước 2. *Xác định tập láng giềng cho sản phẩm cần tư vấn*. Tại bước này ta chỉ cần sắp xếp các giá trị p_{xy} theo thứ tự giảm dần. Sau đó chọn tập K sản phẩm đầu tiên làm tập láng giềng của sản phẩm x [16]. Ký hiệu tập láng giềng của sản phẩm $x \in P$ là K_x.

Bước 3. *Dự đoán quan điểm của người dùng đối với các sản phẩm mới*. Phương pháp phổ biến để sinh ra dự đoán quan điểm của người dùng $i \in U$ cho sản phẩm mới $x \in P$ theo công thức (35)[16].

$$r_{ix} = \frac{\sum_{y \in K_x} p_{xy} r_{iy}}{\sum_{y \in K_x} |p_{xy}|} \quad (35)$$

Bước 4. *Chọn K sản phẩm mới có r_{ix} cao nhất tư vấn cho người dùng i*.

4.3. Độ tương tự giữa các cặp người dùng dựa trên đồ thị

Phương pháp Hybrid-UserBased được đề xuất trong Mục 3.1 có thể thực hiện dễ dàng trên đồ thị bằng cách xem xét tất cả các đường đi độ dài 2 từ đỉnh người dùng đến đỉnh người dùng trên đồ thị [5, 7]. Ví dụ để xác định mức độ tương tự giữa người dùng u_1 và u_2 trên đồ thị trong Hình 4 ta dựa vào các đường đi : u_1-p_1-u_2, u_1-c_1-u_2, u_1-c_3-u_2. Trọng số của mỗi đường đi được tính bằng tích của trọng số các cạnh. Tổng trọng số tất cả các đường đi từ đỉnh $i \in U$ đến đỉnh $j \in U$ chính là độ tương tự giữa hai người dùng này. K người dùng có tổng trọng số các đường đi từ đỉnh $i \in U$ đến đỉnh $j \in U$ có trọng số lớn nhất chính là tập láng giềng của người dùng i. Sau đó sử dụng tập láng giềng để sinh ra dự đoán cho người dùng i.

Một trong những thách thức lớn nhất của hệ tư vấn là vấn đề dữ liệu thưa [1, 3]. Vấn đề dữ liệu thưa xảy ra khi các giá trị đánh giá $r_{ix} \neq 0$ rất ít (dưới 1%). Số lượng các cạnh (i, x) thấp làm cho việc xác định các cạnh (i, s) cũng thấp. Điều này làm cho kết quả dự đoán của các phương pháp trên đạt kết quả không cao. Để hạn chế điều này, chúng tôi tiến hành mở rộng độ dài đường đi từ đỉnh người dùng đến đỉnh người dùng để tận dụng mối liên hệ gián tiếp giữa các cặp người dùng và các cặp đặc trưng nội dung khác nhau. Các đường đi có thể là các cạnh đánh giá (i, x), cạnh (i, s), cạnh (q, x) hoặc cạnh (q, s). Ví dụ để xác định mức độ tương tự giữa u_2 và u_3 với đồ thị biểu diễn bài toán tư vấn kết hợp trong Hình 4, ta có thể sử dụng các đường đi u_2-p_1-u_1-p_3-u_3, u_2-p_4-t_3-p_2-u_3, u_2-c_1-t_4-p_3-u_3. Điều này là hoàn toàn hợp lý vì u_2 thích p_1, p_1 được u_1 thích, u_1 thích p_3 , p_3 được u_3 thích nên gián tiếp u_2 tương tự với u_3 ở một mức độ nào đó. Hoặc u_2 thích p_4, p_4 được người dùng có đặc trưng t_3 thích, người dùng có đặc trưng t_3 thích p_2, u_3 thích t_2 nên gián tiếp u_2 tương tự với u_3 ở một mức độ nào đó. Hoặc u_2 thích đặc trưng c_1, c_1 phù hợp với tập người dùng có đặc trưng t_4, t_4 phù hợp với sản phẩm p_3, u_3 thích p_3 nên cũng gián tiếp u_2 tương tự với u_3 ở một mức độ nào đó.

Vì đồ thị tư vấn kết hợp là đồ thị hai phía nên các đường đi từ đỉnh người dùng đến đỉnh người dùng luôn có độ dài chẵn (2, 4, 6, 8) [7]. Trọng số của mỗi đường đi được tính bằng tích các trọng số các cạnh nên đường đi qua các cạnh có trọng số cao vẫn được đánh giá cao, đường đi qua các cạnh có trọng số thập vẫn được đánh giá thấp. Để ưu tiên cho các đường đi ngắn (độ dài 2) chúng tôi sử dụng tham số α ($0<\alpha<1$) để đánh thấp trọng số các đường đi có độ dài cao. Cụ thể, phương pháp ước lượng tổng trọng số các đường đi độ dài L từ đỉnh người dùng đến đỉnh người dùng được xác định theo công thức (36) [7].

$$R^L = \begin{cases} R.R^T & nếu\ L = 2 \\ \alpha.R.R^T.R^{L-2} & nếu\ L = 4, 6, 8, \ldots \end{cases} \quad (36)$$

Trong đó, L là độ dài đường đi, R là ma trận đánh giá mở rộng được xác định theo (21), R^T là ma trận chuyển vị của R. Giá trị L chẵn được xác định khi tất cả $r_{ij}^L \neq 0$ [7]. Tổng trọng số các đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $j \in U$ là mức độ tương tự giữa hai người dùng này. K người dùng $j \in U$ có r_{ij}^L lớn nhất chính là tập láng giềng của người dùng $i \in U$. Dựa trên nhận xét này, chúng tôi điều chỉnh Bước 1 của thuật toán Hybrid-UserBased trong Mục 3.1 thành thuật toán Hybrid-UserBased-Graph trong Hình 5.

Thuật toán Hybrid-UserBased-Graph:

Đầu vào:

- *Ma trận đánh giá mở rộng R*= (r_{ix}) *biểu diễn đồ thị kết hợp được xác định theo* (21).
- *i*∈*U là người dùng cần được tư vấn.*
- *K là số lượng người dùng của tập láng giềng.*

Đầu ra:

- *Dự đoán x:* r_{ix}| $x \in P \backslash P_i$(*quan điểm của người dùng i đối với các sản phẩm mới* $x \in P$).

Các bước tiến hành:

Bước 1. *Tính toán mức độ tương tự giữa các cặp người dùng trên đồ thị kết hợp*:

L ⇐*2;//Thiết lập độ dài đường đi ban đầu L=2*

Repeat

$$R^L = \begin{cases} R.R^T & nếu\ L = 2 \\ \alpha.R.R^T.R^{L-2} & nếu\ L = 4, 6, 8, \ldots \end{cases}$$

L⇐*L + 2*; *//Tăng độ dài đường đi.*

Until ($r_{ij}^L \neq 0$với mọi $j \in (U \setminus i)$);

Bước 2. *Xác định tập láng giềng cho người dùng i*∈*U.*

- Sắp xếp $r_{ij}^L \neq 0$ theo thứ tự giảm dần (*i*≠*j*).
- Chọn *K* người dùng *j*∈ *U* đầu tiên làm tập láng giềng của người dùng *i* (Ký hiệu tập láng giềng của người dùng *i*∈ *U* là K_i).

Bước 3. *Dự đoán quan điểm của người dùng i đối với các sản phẩm*$x \in P \backslash P_i$.

$r_{ix} = \frac{1}{|K_i|} \sum_{j \in K_i} r_{ix}$;

Bước 4. *Chọn K sản phẩm có* r_{ix}*cao nhất tư vấn cho người dùng i.*

Hình 5. Thuật toán Hybrid-UserBased-Graph

4.4. Độ tương tự giữa các cặp sản phẩm dựa trên đồ thị

Phương pháp Hybrid-ItemBased được đề xuất trong Mục 3.2 cũng có thể thực hiện dễ dàng trên đồ thị bằng cách xem xét tất cả các đường đi độ dài 2 từ đỉnh sản phẩm đến đỉnh sản phẩm trên đồ thị [7]. Ví dụ để xác định mức độ tương tự giữa sản phẩm p_1 và p_3 trên đồ thị trong Hình 4 ta dựa vào các đường đi: p_1-u_1-p_3, p_1-t_1-p_3, p_1-t_2-p_3. Trọng số của mỗi đường đi được tính bằng tích của trọng số các cạnh. Tổng trọng số tất cả các đường đi từ đỉnh $x \in P$ đến đỉnh $y \in P$ chính là độ tương tự giữa hai sản phẩm. *K* sản phẩm có tổng trọng số các đường đi từ đỉnh $x \in P$ đến đỉnh $y \in P$ có trọng số lớn nhất chính là tập láng giềng của sản phẩm *x*. Sau đó sử dụng tập láng giềng của sản phẩm để dự đoán các sản phẩm phù hợp nhất đối với người dùng *i* [7].

Để hạn chế ảnh hưởng của vấn đề dữ liệu thưa, chúng tôi tiến hành mở rộng độ dài đường đi từ sản phẩm đến đỉnh sản phẩm để tận dụng mối liên hệ gián tiếp giữa các cặp sản phẩm và các cặp đặc trưng nội dung. Các đường đi có thể là các cạnh đánh giá (*i*, *x*), cạnh (*i*, *s*), cạnh (*q*, *x*) hoặc cạnh (*q*, *s*). Ví dụ để xác định mức độ tương tự giữa p_1 và p_2 với đồ thị biểu diễn bài toán tư vấn kết hợp trong Hình 4, ta có thể sử dụng các đường đi p_1-u_1-p_3-u_2-p_2, p_1-u_2-p_4-t_1-p_2, p_1-t_2-c_3-u_3-p_2. Tính hợp lý của phép suy diễn này cũng được lý giải tương tự như trường hợp tính toán mức độ tương tự giữa các cặp người dùng.

Vì đồ thị tư vấn kết hợp là đồ thị hai phía nên các đường đi từ đỉnh sản phẩm đến đỉnh sản phẩm luôn có độ dài chẵn (2, 4, 6, 8) [5, 7]. Trọng số của mỗi đường đi được tính bằng tích các trọng số các cạnh nên đường đi qua các cạnh có trọng số cao vẫn được đánh giá cao, đường đi qua các cạnh có trọng số thấp vẫn được đánh giá thấp. Để ưu tiên cho các đường đi ngắn (độ dài 2) chúng tôi sử dụng tham số α (0<α<1) để đánh thấp trọng số các đường đi có độ dài cao. Cụ thể, phương pháp ước lượng tổng trọng số các đường đi độ dài *L* từ đỉnh sản phẩm đến đến đỉnh sản phẩm được xác định theo công thức (37) [7].

$$R^L = \begin{cases} R^T.R & nếu\ L = 2 \\ \alpha.R^T.R\ .R^{L-2} & nếu\ L = 4, 6, 8, \ldots \end{cases} \quad (37)$$

Trong đó, *L* là độ dài đường đi, *R* là ma trận đánh giá mở rộng được xác định theo (20), R^T là ma trận chuyển vị của *R*. Giá trị *L* chẵn được xác định khi tất cả $r_{xy}^L \neq 0$ [7]. Tổng trọng số các đường đi độ dài *L* từ đỉnh $x \in P$ đến đỉnh $y \in P$ là mức độ tương tự giữa hai người dùng này. *K* sản phẩm $y \in P$ có r_{xy}^L lớn nhất chính là tập láng giềng của sản phẩm $x \in P$. Dựa trên nhận xét này, chúng tôi điều chỉnh Bước 1 của thuật toán Hybrid-ItemBased trong Mục 3.2 thành thuật toán Hybrid-ItemBased-Graph trong Hình 6.

Thuật toán Hybrid-ItemBased-Graph:

Đầu vào:

- *Ma trận đánh giá mở rộng R= (r_{ix}) biểu diễn đồ thị kết hợp được xác định theo* (21).
- *i∈U là người dùng cần được tư vấn.*
- *K là số lượng sản phẩm của tập láng giềng.*

Đầu ra:

- *Dự đoán x: r_{ix}| x∈P\P_i (quan điểm của người dùng i đối với các sản phẩm mới x∈P).*

Các bước tiến hành:

Bước 1. *Tính toán mức độ tương tự giữa các cặp sản phẩm trên đồ thị kết hợp*:

L ⇐2;//Thiết lập độ dài đường đi ban đầu L=2

Repeat

$$R^L = \begin{cases} R^T.R & nếu\ L = 2 \\ \alpha.R^T.R.R^{L-2} & nếu\ L = 4, 6, 8, \dots \end{cases}$$

L⇐L + 2; //Tăng độ dài đường đi.

Until (r_{xy}^L≠0với mọi *y*∈(*P* \ *x*));

Bước 2. *Xác định tập láng giềng cho sản phẩm x∈P.*

- Sắp xếp r_{xy}^L≠0 theo thứ tự giảm dần (*x*≠*y*).
- Chọn *K* sản phẩm *y*∈*P* đầu tiên làm tập láng giềng của sản phẩm *x* (Ký hiệu tập láng giềng của sản phẩm *x*∈*P* là K_x).

Bước 3. *Dự đoán mức độ phù hợp của người dùng i đối với các sản phẩm x∈P\P_i.*

$r_{ix} = \frac{1}{|K_x|}\sum_{x\in K_x} r_{ix}$;

Bước 4. *Chọn K sản phẩm có r_{ix} cao nhất tư vấn cho người dùng i.*

Hình 6. Thuật toán Hybrid-ItemBased-Graph

V. THỬ NGHIỆM VÀ 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 [24]. Phương pháp trình bày ở trên được đánh giá và so sánh với các phương pháp khác theo thủ tục mô tả dưới đây.

5.1. Dữ liệu thử nghiệm

Thuật toán lọc kết hợp được thử nghiệm trên bộ dữ liệu MovieLens của nhóm nghiên cứu GroupLens thuộc Trường Đại học Minnesota [24]. Bộ dữ liệu gồm *100.000* đánh giá của *943* người dùng cho *1682* phim. Giá trị đánh giá được thực hiện từ *1* đến *5*. Mức độ thưa thớt dữ liệu đánh giá là *99.1%*. Tập đặc trưng nội dung sản phẩm được chọn là 18 thể loại phim khác nhau [18]. Tập đặc trưng nội dung người dùng cũng được cung cấp kèm theo tập đặc trưng nội dung phim [24]. Chọn giá trị θ = 15 theo phương pháp được mô tả ở trên để xác định w_{is}, v_{qx}, d_{qs} theo công thức (8), (12), và (16) theo thứ tự. Chọn α=0.8 để xác định trọng số đường đi cho các công thức (36), (37).

5.2. 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 75% đánh giá và tập U_{te} chứa 25% 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[8, 14].

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.

$$MAE_u = \frac{1}{|P_u|}\sum_{y\in P_u} |\hat{r}_{uy} - r_{uy}| \quad (38)$$

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 [8, 14].

$$MAE = \frac{\sum_{u\in U_{te}} MAE_u}{|U_{te}|} \quad (39)$$

5.3. So sánh và đánh giá

Các phương pháp tư vấn kết hợp *Hybrid-UserBased, Hybrid-ItemBased, Hybrid-UserBased-Graph, Hybrid-ItemBased-Graph* được trình bày trong mục 3.1, 3.2, 3.3 và 3.4 được so sánh với những phương pháp lọc cộng tác sau:

- Phương pháp CF-UserBased sử dụng độ tương quan Pearson. Đây là phương pháp tư vấn cộng tác chuẩn dựa vào người dùng chỉ thực hiện dự đoán dựa trên tập giá trị đánh người dùng[15].

- Phương pháp CF-ItemBased sử dụng độ tương quan Pearson. Đây là phương pháp lọc cộng tác chuẩn dựa vào sản phẩm chỉ thực hiện dự đoán dựa trên tập giá trị đánh người dùng[16].

Phương pháp thử nghiệm được thực hiện chọn ngẫu nhiên *300, 600, 900* người dùng trong tập MovieLens làm dữ liệu huấn luyện. Chọn ngẫu nhiên 100, 200, 300 người dùng trong số còn lại để làm tập kiểm tra. Giá trị *MAE* trong Bảng 7 và Bảng 8 được ước lượng từ trung bình của 10 lần thử nghiệm ngẫu nhiên.

Kết quả trong Bảng 7 cho thấy phương pháp lọc dựa vào người dùng thuần túy CF-UserBased cho lại giá trị MAE lớn nhất so với các phương pháp còn lại. Điều này có thể lý giải hạn chế của phương pháp lọc cộng tác khi quá trình huấn luyện chỉ dựa vào tập rất nhỏ các giá trị $r_{ix}\neq 0$. Khi kích thước tập dữ liệu huấn luyện lớn kết quả dự đoán phương pháp dần được cải thiện. Cụ thể giá trị MAE trên tập dữ liệu gồm 300, 600, 900 người dùng lần lượt là (0.856, 0.834, 0.838), (0.794, 0.775, 0.764), (0.778, 0.745, 0.748) theo thứ tự. Kích cỡ tập láng giềng lớn thực hiện thiếu ổn định. Tập láng giềng lớn không tỉ lệ thuận với kết quả dự đoán. Kết quả này hoàn toàn phù hợp với những nghiên cứu trước đây.

Phương pháp Hybrid-UserBased cho lại giá trị MAE thấp hơn nhiều so với phương pháp CF-UserBased. Cụ thể với tập láng giềng K=10 và kích thước tập dữ liệu huấn luyện gồm 300, 600, 900 người dùng thì phương pháp cho lại giá trị MAE lần lượt là 0.642, 0.611, 0.607 so với 0.856, 0.794, 0.788 của phương pháp CF-UserBased; với K=20 giá trị MAE lần lượt là 0.586, 0.597, 0.611 so với 0.834, 0.775, 0.745 của phương pháp CF-UserBased;; với K=30 giá trị MAE lần lượt là 0.513, 0.553, 0.609 so với 0.838, 0.764, 0.748 của phương pháp CF-UserBased. Số lượng người dùng trong tập láng giềng lớn kết quả dự đoán cũng thực hiện ổn định hơn. Điều này có thể giải thích phương pháp Hybrid-UserBased tính toán được mức độ tương tự giữa các cặp người dùng chính xác hơn vì phương pháp được thực hiện trên toàn bộ tập dữ liệu đánh giá và hồ sơ người dùng. Chính vì lý do đó phương pháp xác định được tập láng giềng của người dùng hiện thời tốt hơn để đưa ra kết quả dự đoán.

Phương pháp Hybrid-UserBased-Graph cho lại giá trị MAE thấp nhất so với tất cả các phương pháp còn lại. Giá trị MAE trên các tập dữ liệu huấn luyện và tập láng giềng dao động trong khoảng $0.515\leq MAE\leq 0.548$. Phương pháp cho lại kết quả khá ổn định trên các tập tập dữ liệu và tập láng giềng có kích cỡ khác nhau. Điều này có thể khẳng định phương pháp xác định độ tương tự dựa trên tập các đường đi từ đỉnh người dùng đến đỉnh người dùng là hoàn toàn tin cậy. Độ tương tự giữa các cặp người dùng trên đồ thị đã tích hợp được tất cả các mối quan hệ giữa người dùng và sản phẩm, người dùng và đặc trưng nội dung sản phẩm, sản phẩm và đặc trưng người dùng, đặc trưng sản phẩm và đặc trưng người dùng.

Bảng 7. Giá trị MAE của các phương pháp tư vấn dựa vào người dùng

Kích thước tập dữ liệu huấn luyện	Phương pháp	Kích thước của tập láng giềng		
		10	20	30
300 người dùng	CF-UserBased	0.856	0.834	0.838
	Hybrid-UserBased	**0.642**	**0.586**	**0.513**
	Hybrid-UserBased-Graph	**0.521**	**0.518**	**0.512**
600 người dùng	Hybrid-CF-UserBased	0.794	0.775	0.764
	Hybrid-UserBased	**0.611**	**0.597**	**0.553**
	Hybrid-UserBased-Graph	**0.515**	**0.515**	**0.537**
900 người dùng	CF-UserBased	0.788	0.745	0.748
	Hybrid-UserBased	**0.607**	**0.611**	**0.609**
	Hybrid-UserBased-Graph	**0.548**	**0.519**	**0.517**

Giá trị MAE trong Bảng 8 của các phương pháp lọc dựa vào sản phẩm cũng có diễn biến tương tự như phương pháp lọc dựa vào người dùng. Giá trị MAE của phương pháp lọc kết hợp Hybrid-ItemBased nhỏ hơn nhiều so với phương pháp CF-ItemBased. Lý do để điều này xảy ra chỉ có thể lý giải phương pháp tính toán mức độtương tựgiữa các cặp sản phẩm được thực hiện trên tập giá trị đánh giá và hồ sơ sản phẩm thực hiện chính xác hơn các phép đo chỉ dựa vào tập giá trị đánh giá. Giá trị MAE của phương pháp Hybrid-ItemBased-Graph thấp hơn đáng kể so với phương pháp Hybrid-ItemBased. Điều này cũng chỉ có thể giải thích độ tương tự giữa các sản phẩm dựa trên đồ thị đã kết hợp được tất cả các mối quan hệ gián tiếp giữa người dùng, sản phẩm, hồ sơ người dùng và hồ sơ sản phẩm.

Bảng 8. Giá trị MAE của các phương pháp tư vấn dựa vào sản phẩm

Kích thước tập dữ liệu huấn luyện	Phương pháp	Kích thước của tập láng giềng		
		5	10	20
300 người dùng	CF-ItemBased	0.843	0.837	0.835
	Hybrid-ItemBased	**0.622**	**0.622**	**0.607**
	Hybrid-ItemBased -Graph	**0.612**	**0.589**	**0.571**
600 người dùng	CF-ItemBased	0.814	0.816	0.844
	Hybrid-ItemBased	**0.651**	**0.637**	**0.613**
	Hibrid-ItemBased -Graph	**0.537**	**0.525**	**0.517**
900 người dùng	CF-ItemBased	0.793	0.786	0.743
	Hybrid-ItemBased	**0.568**	**0.587**	**0.543**
	Hybrid-ItemBased -Graph	**0.548**	**0.519**	**0.511**

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất một mô hình hợp nhất giữa phương pháp tư vấn cộng tác và tư vấn the nội dung. Mô hình được xây dựng bằng cách dịch chuyển bài toán tư vấn kết hợp về bài toán tư vấn cộng tác thuần túy để tận dụng những ưu điểm của phương pháp này. Phương pháp dịch chuyển được tiến hành bằng cách xây dựng hồ sơ người dùng của lọc nội dung dựa trên đánh giá tự nhiên của người dùng đối với các sản phẩm. Sau đó, thiết lập mối quan hệ trực tiếp giữa người dùng và từng đặc trưng nội dung sản phẩm. Bằng cách này ta mở rộng được ma trận đánh giá của lọc cộng tác về phía các sản phẩm. Tiếp đến, quá trình xây dựng hồ sơ sản phẩm cũng được thực hiện trên thói quen sử dụng sản phẩm một cách tự nhiên của người dùng đối với các sản phẩm. Dựa trên hồ sơ sản phẩm, chúng tôi thiết lập mối quan hệ trực tiếp giữa sản phẩm và mỗi đặc trưng nội dung người dùng. Bằng cách này ta mở rộng được ma trận đánh giá của lọc cộng tác về phía người dùng. Cuối cùng, chúng tôi tìm cách xác định mối quan hệ tiềm ẩn giữa mỗi đặc trưng người dùng với các đặc trưng sản phẩm dựa trên hồ sơ người dùng và hồ sơ sản phẩm. Mô hình cuối cùng nhận được là mở rộng của mô hình tư vấn cộng tác cơ bản.

Sau khi thu gọn về bài toán lọc cộng tác, ma trận đánh giá mở rộng được đề xuất trong bài báo đã tích hợp đầy đủ các giá trị đánh giá của lọc cộng tác, hồ sơ người dùng, hồ sơ sản phẩm, mối liên hệ giữa hồ sơ người dùng và hồ sơ nội dung sản phẩm của lọc nội dung. Trọng số các đặc trưng nội dung trong hồ sơ người dùng, hồ sơ sản phẩm và mối liên hệ giữa các đặc trưng nội dung có cùng metric với giá trị đánh giá. Chính vì vậy, các phương pháp tư vấn cộng tác dựa vào bộ nhớ hoặc các phương pháp tư vấn cộng tác dựa trên mô hình đều có thể triển khai trên ma trận đánh giá mở rộng. Để tận dụng ưu thế của mô hình đồ thị, chúng tôi đề xuất xây dựng một độ tương quan khai thác gián tiếp các mối quan hệ giữa người dùng, sản phẩm, đặc trưng người dùng, đặc trưng sản phẩm để nâng cao kết quả dự đoán. Kết quả thử nghiệm các phương pháp tư vấn kết hợp đề xuất cho thấy chất lượng tư vấn được cải thiện đáng kể so với các phương pháp tư vấn cơ bản. Chúng tôi tin tưởng rằng, mô hình cũng sẽ cho lại kết quả tốt đối với các phương pháp tư vấn dựa trên mô hình. Những kết quả này sẽ được trình bày trong những nghiên cứu tiếp theo của bài báo.

VII. 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. Adomavicius G., Tuzhilin A., "*Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*", IEEE Transactions On Knowledge And Data Engineering, vol. 17, No. 6, 2005.
3. Robin D. Burke, "*Hybrid Recommender Systems: Survey and Experiments*". User Model. User-Adapt. Interact. 12(4): 331-370 (2002).
4. 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.
5. Nguyen Duy Phuong, Le Quang Thang, Tu Minh Phuong, *"A Graph-Based Method for Combining Collaborative and Content-Based Filtering.* PRICAI 2008: 859-869.
6. Nguyen Duy Phuong, Tu Minh Phuong, "Collaborative Filtering by Multi-task Learning", RIVF 2008, pp: 227-232.
7. Do Thi Lien, Nguyen Duy Phuong, "*Collaborative Filtering with a Graph-based Similarity Measure".* ComManTel, 2014, pp. 251-256.
8. Asela Gunawardana, Guy Shani, "*A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*. Journal of Machine Learning Research 10: 2935-2962 (2009).
9. Asela Gunawardana, Christopher Meek, " *A unified approach to building hybrid recommender systems*". RecSys 2009: 117-124.
10. Robin D. Burke, Fatemeh Vahedian, Bamshad Mobasher, "*Hybrid Recommendation in Heterogeneous Networks*". UMAP 2014: 49-60.
11. 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.
12. 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).
13. Pazzani, M. J. "*A framework for collaborative, content-based and demographic filtering*", Artificial Intelligence Review 13(5-6), 393–408 (1999).
14. 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.
15. 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).
16. Sarwar B., Karypis G., Konstan J., and Riedl J., "*Item-Based Collaborative Filtering Recommendation Algorithms*", Proc. 10th Int'l WWW Conf (2001).

17. 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).
18. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. : Combining contentbased and collaborative fillters in an online newspaper. Proceedings of ACM SIGIR Workshop on Recommender Systems.(1999).
19. Basu, C., Hirsh, H., And Cohen, W.: Recommendation as classification: Using social and content-based information in recommendation. In Proceedings of the 15th National Conference on Artificial Intelligence, 714–720. (1998).
20. Popescul A., Ungar L.H., Pennock D.M., and Lawrence S.: Probabilistic Models for Unified Collaborative and Content-Based Eecommendation in Sparse-Data Environments, Proc. 17th Conf. Uncertainty in Artificial Intelligence, (2001).
21. Balisico J., Hofman T.: Unifying collaborative and content-based filtering. In Proceedings. of Int. Conf. on Machine learning (ICML-04) (2004).
22. Crammer, K., and Singer, Y: Pranking with ranking. Advances in Neural Information Processing Systems 14 pp. 641-647. (2002).
23. Aggarwal C.C., Wolf J.L., Wu K.L., and Yu P.S.: Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering, Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. (1999).
24. http://www.grouplens.org/.

A GRAPH-BASED MODEL FOR HIBRID RECOMMENDER SYSTEM

Do Thi Lien, Nguyen Xuan Anh, Nguyen Duy Phuong, Tu Minh Phuong

Abstract - *Recommender systems are the capable systems of providing essentialinformation and removing unessentialinformation for Internet users. The recommender systems are built based on two main information filtering techniques: Collaborative filtering and content-based filtering. Each method exploits particular aspects related to content features or product usage habit of users in the past to predict a brief list of the most suitable products with each user. In this paper, we propose a new unify method between collaborative filtering recommendation and content-based filtering recommendation based on graph model. The model allows us to shift hybrid filtering recommender problem to collaborative filtering recommender problem, then build new similar measures based on graph to determine similarities between two users or two items, these similar measures are used to predict suitable products for users in the system. The experimental results on real data sets show that the proposed methods achieve superior performance compared to baseline methods.*

Keywords - *Collaborative Filtering Recommendation*, *Content-based Filtering Recommendation*, *Hybrid Filtering Recommendation System*, *Item-Based Recommendation*, *User-Based Recommendation*.