Một phương pháp lọc cộng tác dựa trên mô hình đồ thị hai phía

A Collaborative Filtering Method Based on Bipartite Graph Model

Mai Thị Như và Nguyễn Duy Phương

Abstract: *Collaborative filtering is a technique to predict the utility of items for a particular user by exploiting the behavior patterns of a group of users with similar preferences. This method has been widely successful in many e-commerce systems. In this paper, we present an effective collaborative filtering method based on general bipartite graph representation. The weighted bipartite graph representation is suitable for all of the real current data sets of collaborative filtering. The prediction method is solved by the basic search problem on the graph that can be easy to implement for the real applications. Specially, the model tackled the effect of the sparsity problem of collaborative filtering by expanding search length from the user node to the item node. By this way, some users or items can not be detemined by the correlations but can be computed by the graph model. Experimental results on the real data sets show that the proposed method improve significantly prediction quality for collaborative filtering.*

I. PHÁT BIỂU BÀI TOÁN LỌC CỘNG TÁC

Cho tập hợp hữu hạn $U = \{u_1, u_2,\ldots, u_N\}$ là tập gồm N người dùng, $P = \{p_1, p_2,\ldots, p_M\}$ là tập gồm M sản phẩm. Mỗi sản phẩm $p_x \in P$ có thể là hàng hóa, phim, ảnh, tạp chí, tài liệu, sách, báo, dịch vụ hoặc bất kỳ dạng thông tin nào mà người dùng cần đến. Để thuận tiện trong trình bày, ta viết $p_x \in P$ ngắn gọn thành $x \in P$; và $u_i \in U$ là $i \in U$.

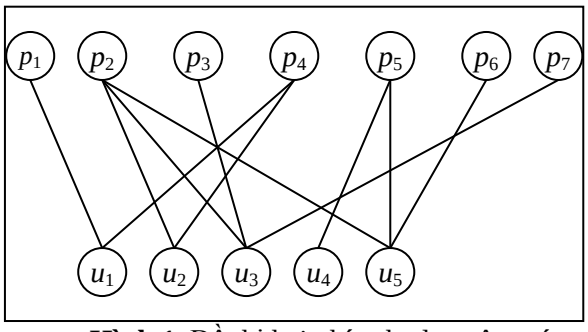
Mối quan hệ giữa tập người dùng U và tập sản phẩm P được biểu diễn thông qua ma trận đánh giá $R = (r_{ix})$, $i = 1...N$, $x = 1...M$. Mỗi giá trị r_{ix} biểu diễn đánh giá của người dùng $i \in U$ cho sản phẩm $x \in P$. Giá trị r_{ix} có thể được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị $r_{ix} = \varnothing$ được hiểu người dùng i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm x.

Tiếp đến ta ký hiệu, $P_i \subseteq P$ là tập các sản phẩm được đánh giá bởi người dùng $i \in U$ và $U_x \subseteq U$ là tập các người dùng đã đánh giá sản phẩm $x \in P$. Với một người dùng cần được tư vấn $a \in U$ (được gọi là người dùng hiện thời, hay người dùng tích cực), bài toán lọc cộng tác là dự đoán đánh giá của người dùng a đối với những mặt hàng $x \in (P \backslash P_a)$, trên cơ sở đó tư vấn cho người dùng a những sản phẩm được đánh giá cao.

Bảng 1 thể hiện một ví dụ với ma trận đánh giá $R = (r_{ij})$ trong hệ gồm 5 người dùng $U = \{u_1, u_2, u_3, u_4, u_5\}$ và 7 sản phẩm $P = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7,\}$. Mỗi người dùng đều đưa ra các đánh giá của mình về các sản phẩm theo thang bậc {1,2,3,4,5}. Đối với tập dữ liệu MovieLens [11], $r_{ix} = 5$ được hiểu là người dùng i đánh giá phim x ở mức độ "*rất tốt*"; $r_{ix} = 4$ được hiểu là người dùng i đánh giá "*tốt*"; $r_{ix} = 3$ được hiểu là người dùng i đánh giá phim x ở mức độ "*bình thường*"; $r_{ix} = 2$ được hiểu là người dùng i đánh giá phim x ở mức độ "*kém*"; $r_{ix} = 1$ được hiểu là người dùng i đánh giá phim x ở mức độ "*rất kém*". Giá trị $r_{ij}=\varnothing$ được hiểu là người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j. Các ô được đánh dấu '?' thể hiện giá trị hệ thống cần dự đoán cho người dùng u_5.

Bảng 1. Ma trận đánh giá của lọc cộng tác.

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	∅	1	5	∅	1	∅
u_2	∅	5	2	5	1	∅	2
u_3	2	4	5	1	∅	∅	4
u_4	1	2	∅	∅	5	2	∅
u_5	?	4	?	1	4	5	?



Hình 1. Đồ thị hai phía cho lọc cộng tác

II. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Có hai hướng tiếp cận chính giải quyết bài toán lọc cộng tác bằng mô hình đồ thị: Lọc cộng tác dựa trên mô hình đồ thị tổng quát và Lọc cộng tác dựa trên mô hình đồ thị hai phía [3,4,6,7]. Để thuận tiện cho việc trình bày mô hình đề xuất, chúng tôi tóm tắt lại những nghiên cứu về mô hình đồ thị hai phía cho lọc cộng tác của Huang và các cộng sự [3,4].

Trong mô hình này, Huang xem xét bài toán lọc cộng tác như bài toán tìm kiếm trên đồ thị hai phía, một phía là tập người dùng *U*, phía còn lại là tập sản phẩm *P*. Cạnh nối giữa người dùng $i \in U$ đến sản phẩm $x \in P$ được thiết lập nếu người dùng *i* đánh giá "*tốt*" hoặc "*rất tốt*" sản phẩm *x*. Ví dụ với ma trận đánh giá được cho trong Bảng 1, các giá trị đánh giá r_{ix} =4, r_{ix} = 5 sẽ được biến đổi thành 1, những giá trị còn lại được biến đổi thành 0. Khi đó, ma trận kề biểu diễn đồ thị hai phía được thể hiện trong Bảng 2, đồ thị hai phía tương ứng theo biểu diễn được thể hiện trong Hình 1.

Bảng 2. Ma trận kề biểu diễn đồ thị hai phía.

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	0	0	1	0	0	0
u_2	0	1	0	1	0	0	0
u_3	0	1	1	0	0	0	1
u_4	0	0	0	0	1	0	0
u_5	0	1	0	0	1	1	0

Phương pháp dự đoán trên đồ thị được thực hiện bằng thuật toán lan truyền mạng để tìm ra số lượng đường đi độ dài *L* từ đỉnh người dùng $i \in U$ đến đỉnh sản phẩm $x \in P$. Những sản phẩm $x \in P$ có số lượng đường đi nhiều nhất đến người dùng $i \in U$ sẽ được dùng để tư vấn cho người dùng này [3].

Với phương pháp biểu diễn và dự đoán nêu trên, chúng tôi đã tiến hành kiểm nghiệm trên các bộ dữ liệu thực và nhận thấy một số những hạn chế dưới đây.

Thứ nhất, biểu diễn của Huang chỉ quan tâm đến các giá trị đánh giá "*tốt*" hoặc "*rất tốt*" và bỏ qua các giá trị đánh giá "*kém*" hoặc "*rất kém*". Đối với các hệ thống lọc cộng tác thực tế, mức đánh giá của người dùng được chia thành nhiều thang bậc khác nhau (tập dữ liệu MovieLens có 5 mức đánh giá, tập BookCrossing có 10 mức đánh giá) [11,12]. Chính vì vậy, biểu diễn này chưa thực sự phù hợp với các hệ thống lọc cộng tác hiện nay. Mặt khác, các phương pháp dự đoán của lọc cộng tác được thực hiện dựa trên thói quen sử dụng sản phẩm của cộng đồng người dùng có cùng sở thích, do vậy các giá trị đánh giá "*tốt*" hay "*không tốt*" đều phản ánh thói quen sử dụng sản phẩm của người dùng. Việc bỏ qua các giá trị "*không tốt*" sẽ ảnh hưởng rất nhiều đến chất lượng dự đoán thói quen sử dụng sản phẩm của người dùng.

Thứ hai, đối với các hệ thống lọc cộng tác số lượng giá trị đánh giá r_{ix}=∅ nhiều hơn rất nhiều lần số lượng giá trị đánh giá $r_{ix} \neq \emptyset$. Vì vậy, việc bỏ qua các giá trị "*không tốt*" khiến cho vấn đề dữ liệu thưa của lọc cộng tác trở nên trầm trọng hơn. Điều này có thể

thấy rõ trong Bảng 2, các giá trị đánh giá $r_{ix} \leq 3$ được biến đổi thành 0 đã bỏ đi một lượng đáng kể các nhãn phân loại biết trước trong quá trình huấn luyện.

Cuối cùng, phương pháp dự đoán được thực hiện dựa vào số lượng đường đi có độ dài L từ đỉnh người dùng đến đỉnh sản phẩm. Các đường đi được xem có trọng số giống nhau là 1 chưa phản ánh đúng hiện trạng của các bộ dữ liệu thực (tập dữ liệu MovieLens có 5 mức đánh giá [11], tập dữ liệu BookCrossing có 10 mức đánh giá [12]). Chính vì vậy, mô hình chỉ cho lại kết quả thử nghiệm tốt trên các tập dữ liệu có hai mức đánh giá (0, 1). Đối với các tập dữ liệu có nhiều mức đánh giá, kết quả dự đoán của mô hình sẽ cho độ chính xác không cao. Tóm lại, mô hình do Huang đề xuất chỉ phù hợp với các tập dữ liệu về sách có hai mức đánh giá "*tốt*" hoặc "*không tốt*".

Để khắc phục được những hạn chế nêu trên, trong mục tiếp theo chúng tôi đề xuất một mô hình đồ thị hai phía tổng quát cho lọc cộng tác. Trong đó, phương pháp biểu diễn được thực hiện trên đồ thị trọng số phù hợp với tất cả bộ dữ liệu thử nghiệm cho lọc cộng tác. Phương pháp dự đoán được thực hiện dựa trên việc tính toán trọng số của tất cả các đường đi từ đỉnh người dùng đến đỉnh sản phẩm cho phép ta cải thiện được chất lượng dự đoán.

III. MÔ HÌNH ĐỒ THỊ HAI PHÍA ĐỀ XUẤT

Mô hình đồ thị hai phía có trọng số đề xuất mở rộng phương pháp tiếp cận trong [1,3,4] ở hai điểm chính: Phương pháp biểu diễn đồ thị và phương pháp dự đoán trên đồ thị. Phương pháp được tiến hành như sau.

III.1. Phương pháp biểu diễn đồ thị

Không hạn chế tính tổng quát của bài toán, ta có thể giả sử $r_{ix} = +v$ nếu người dùng i đánh giá sản phẩm x "*tốt*" ở mức độ v, $r_{ix} = -v$ nếu người dùng i đánh giá sản phẩm x "*không tốt*" ở mức độ $-v$, trong đó $v \in [-1,1]$.

$$r_{ix} = \begin{cases} v & \text{Nếu người dùng } i \text{ thích sản phẩm } x \text{ ở mức độ } v. \\ \varnothing & \text{Nếu người dùng } i \text{ chưa biết đến sản phẩm } x. \\ -v & \text{Nếu người dùng } i \text{ không thích sản phẩm } x \text{ ở mức độ } -v. \end{cases} \quad (1)$$

Đối với các tập dữ liệu thực của lọc cộng tác, ta dễ dàng chuyển đổi biểu diễn thành ma trận đánh giá theo công thức (1) bằng cách chọn một giá trị ngưỡng θ. Những giá trị $r_{ix} > \theta$ được dịch chuyển thành các giá trị dương, ngược lại chuyển đổi thành giá trị âm. Ví dụ với ma trận đánh giá được cho trong Bảng 1, chọn θ=3, khi đó các giá trị r_{ix}= 4, 5 biến đổi thành 0.1, 0.2, các giá trị r_{ix} = 2, 1 biến đổi thành -0.1, -0.2, r_{ix}=3 biến đổi thành ∅ như trong Bảng 3.

Với cách chuyển đổi biểu diễn theo công thức (1), vấn đề lọc cộng tác được biểu diễn như một đồ thị hai phía (Ký hiệu là đồ thị G). Một phía là tập người dùng U, phía còn lại là tập các sản phẩm P. Trong đó, cạnh nối giữa đỉnh phía người dùng $i \in U$ với đỉnh phía sản phẩm $x \in P$ được thiết lập nếu $r_{ix} \neq \varnothing$. Những giá trị đánh giá có $r_{ix} > 0$ biểu diễn người dùng $x \in U$ đánh giá sản phẩm $i \in P$ "*tốt*" ở mức độ r_{ix}. Những giá trị đánh giá có $r_{ix} < 0$ biểu diễn người dùng $x \in U$ đánh giá sản phẩm $i \in P$ "*không tốt*" ở mức độ r_{ix}. Trọng số của mỗi cạnh trên đồ thị được xác định theo công thức:

$$w_{ix} = \begin{cases} r_{ix} & \text{nếu } r_{ix} \neq \varnothing \\ 0 & \text{nếu } r_{ix} = \varnothing \end{cases} \quad (2)$$

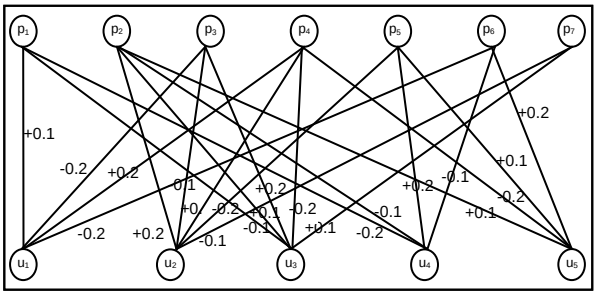
Ví dụ, với hệ lọc cộng tác được cho trong Bảng 3 thì ta sẽ có ma trận trọng số của đồ thị hai phía như trong Bảng 4, đồ thị hai phía tương ứng được thể hiện trong Hình 2.

Bảng 3. Ma trận đánh giá của lọc cộng tác.

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	0.1	∅	-0.2	0.2	∅	-0.2	∅
u_2	∅	0.2	-0.1	0.2	-0.2	∅	-0.1
u_3	-0.1	0.1	0.2	-0.2	∅	∅	0.1
u_4	-0.2	-0.1	∅	∅	0.2	-0.1	∅
u_5	?	0.1	?	-0.2	0.1	0.2	?

Bảng 4. Ma trận trọng số của đồ thị hai phía.

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	0.1	0	-0.2	0.2	0	-0.2	0
u_2	0	0.2	-0.1	0.2	-0.2	0	-0.1
u_3	-0.1	0.1	0.2	-0.2	0	0	0.1
u_4	-0.2	-0.1	0	0	0.2	-0.1	0
u_5	0	0.1	0	-0.2	0.1	0.2	0



Hình 2. Đồ thị hai phía tổng quát cho lọc cộng tác

So với phương pháp biểu diễn trong [1,3,4], mô hình đề xuất giữ lại tất cả các giá trị đánh giá của người dùng đối với sản phẩm. Ví dụ, với tập dữ liệu MovieLens biểu diễn quan điểm của người dùng đối với các phim theo 5 mức đánh giá từ 1 đến 5. Ta có thể dịch chuyển các giá trị $r_{ix} = 5$ thành *0.2*, $r_{ix} = 4$ thành *0.1*, $r_{ix} = 3$ thành *0.0*, $r_{ix}=2$ thành *-0.1*, $r_{ix} = 1$ thành *-0.2*. Với tập dữ liệu BookCrossing biểu diễn quan điểm của người dùng đối với các sản phẩm theo 10 mức đánh giá từ 0.1, 0.2, 0.3, 0.4, 0.5,…, 1.0. Việc chuyển đổi cũng được tiến hành tương tự như trên bằng cách biến đổi các giá trị $r_{ix}>0.5$ (0.6, 0.7, 0.8, 0.9, 1.0) thành các giá trị dương (0.1, 0.2, 0.3, 0.4, 0.5). Các giá trị $r_{ix}\leq 0.5$ (0.5, 0.4, 0.3, 0.2, 0.1) được biến đổi thành các giá trị âm (-0.1, -0.2, -0.3, -0.4, -0.5). Các bộ dữ liệu khác cũng được biến đổi tương tự tùy thuộc vào các mức đánh giá khác nhau của người dùng. Trong mục tiếp theo chúng tôi trình bày về phương pháp dự đoán trên đồ thị hai phía có trọng số.

III. 2. Phương pháp dự đoán

Khác với các cách tiếp cận trong [3,4], mô hình đề xuất xem xét vấn đề lọc cộng tác như bài toán tìm kiếm trên đồ thị hai phía có trọng số. Do vậy, phương pháp dự đoán của mô hình phụ thuộc vào tổng trọng số của các đường đi có độ dài *L* từ đỉnh người dùng đến đỉnh sản phẩm. Những sản phẩm nào có tổng trọng số các đường đi từ đỉnh người dùng đến đỉnh sản phẩm lớn nhất sẽ được dùng để tư vấn cho người dùng hiện thời.

Trong [3,4], Huang xem xét trọng số tất cả các đường đi có độ dài *L* từ đỉnh người dùng đến đỉnh sản phẩm đều giống nhau và có giá trị *1*. Trong mô hình này, chúng tôi xem xét mỗi đường đi độ dài *L* từ đỉnh người dùng đến đỉnh sản phẩm có trọng số khác nhau. Điều này hoàn toàn phù hợp với các bộ dữ liệu thực có nhiều mức đánh giá khác nhau.

Một điểm khác biệt quan trọng với các mô hình đề xuất trong [3,4] là trọng số các đường đi độ dài *L* từ đỉnh người dùng đến đỉnh sản phẩm có thể nhận giá trị dương hoặc giá trị âm. Các đường đi có trọng số dương phản ánh mức độ đánh giá sản phẩm "*tốt*" của người dùng. Các đường đi có trọng số âm phản ánh mức độ đánh giá sản phẩm "*không tốt*" của người dùng.

Để thực hiện ý tưởng nêu trên, chúng tôi tiến hành tách đồ thị hai phía tổng quát ban đầu thành hai đồ thị con: Đồ thị hai phía với các cạnh có trọng số dương (ký hiệu là G^+) và đồ thị hai phía với các cạnh có trọng số âm (ký hiệu là G^-). Ứng với sản phẩm $x\in P$ chưa được người dùng $i\in U$ đánh giá, quá trình ước lượng mức độ "*tốt*" của sản phẩm *x* đối với người dùng *i* được thực hiện trên đồ thị G^+ bằng cách tính tổng trọng số tất cả các đường đi độ dài *L* từ *x* đến *i*. Tương tự như vậy, quá trình ước lượng mức độ "*không tốt*" của sản phẩm *x* đối với người dùng *i* được thực hiện trên đồ thị G^- bằng cách tính tổng trọng số tất cả các đường đi độ dài *L* từ *x* đến *i*. Hai giá trị này được kết hợp lại sẽ cho ta quan điểm chính xác của người dùng *x* đối với sản phẩm *i*.

Gọi $W^+ = (w_{ix}^+)$ là ma trận trọng số biểu diễn đồ thị G^+, $W^- = (w_{ix}^-)$ là ma trận trọng số biểu diễn đồ thị G^- được xác định theo công thức (3), (4).

$$w_{ix}^+ = \begin{cases} w_{ix} & \text{nếu } w_{ix}>0 \\ 0 & \text{nếu } w_{ix}\le 0 \end{cases} \quad (3)$$

$$w_{ix}^- = \begin{cases} w_{ix} & \text{nếu } w_{ix}<0 \\ 0 & \text{nếu } w_{ix}\ge 0 \end{cases} \quad (4)$$

Vì G^+, G^- là đồ thị hai phía nên mỗi đường đi độ dài L từ đỉnh người dùng $i \in U$ đến đỉnh sản phẩm $x \in P$ đều có độ dài lẻ ($L=1,3,5,7,...$). Trọng số đường đi độ dài L trên đồ thị G^+, G^- được tính bằng tích các trọng số w_{ix}^+, w_{ix}^-. Do $0 \le w_{ix}^+ \le 1; -1 \le w_{ix}^- \le 0$ nên đường đi có độ dài L lớn sẽ có trọng số thấp, đường đi độ dài L nhỏ sẽ có trọng số cao. Trọng số đường đi từ đỉnh người dùng đến đỉnh sản phẩm trên đồ thị G^+ luôn là một số dương, trên đồ thị G^- luôn là một số âm. Ví dụ trên Hình 2, đường đi u_5-p_2-u_2-p_4 có trọng số là $w^+_{54} = 0.1 \times 0.2 \times 0.2 = 0.004$; u_5-p_2-u_3-p_7 có trọng số là $w_{57} = 0.1 \times 0.1 \times 0.1 = 0.001$. Ngược lại, đường đi u_1-p_3-u_2-p_5 có trọng số là $w^-_{15} = (-0.2) \times (-0.1) \times (-0.2) = -0.004$; u_1-p_3-u_2-p_7 có trọng số là $w_{17} = (-0.2) \times (-0.1) \times (-0.1) = -0.002$. Bằng cách này, ta có thể phân biệt được mức độ quan trọng của mỗi đường đi khác nhau trong quá trình dự đoán.

Tổng quát, tổng trọng số các đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ trên đồ thị G^+, G^- được xác định theo công thức (5), (6).

$$(w_{ix}^+)^L = \begin{cases} (w_{ix}^+) & \text{nếu } L=1 \\ (w_{ix}^+).(w_{ix}^+)^T.(w_{ix}^+)^{L-2} & \text{nếu } L=3,5,7,... \end{cases} \quad (5)$$

$$(w_{ix}^-)^L = \begin{cases} (w_{ix}^-) & \text{nếu } L=1 \\ (w_{ix}^-).(w_{ix}^-)^T.(w_{ix}^-)^{L-2} & \text{nếu } L=3,5,7,... \end{cases} \quad (6)$$

Ở đây, $(w_{ix}^+)^L$ là tổng trọng số các đường đi độ dài L từ đỉnh người dùng $i \in U$ đến sản phẩm $x \in P$ trên đồ thị G^+, $(w_{ix}^-)^L$ là tổng trọng số các đường đi độ dài L từ đỉnh người dùng $i \in U$ đến sản phẩm $x \in P$ trên đồ thị G^-, $(w_{ix}^+)^T$, $(w_{ix}^-)^T$ là ma trận chuyển vị của (w_{ix}^+) và (w_{ix}^-).

Giá trị $(w_{ix}^+)^L$ có trọng số luôn dương phản ánh mức độ "*tốt*" của sản phẩm x đối với người dùng i suy diễn trên đồ thị G^+. Giá trị $(w_{ix}^-)^L$ có trọng số luôn âm phản ánh mức độ "*không tốt*" của sản phẩm x đối với người dùng i suy diễn trên đồ thị G^-. Sau khi tính toán $(w_{ix}^+)^L$, $(w_{ix}^-)^L$ các giá trị này sẽ được tổ hợp lại theo công thức (7) để đưa ra kết quả dự đoán quan điểm của người dùng i đối với sản phẩm x.

$$w_{ix}^L = \alpha.(w_{ix}^-)^L + (1-\alpha).(w_{ix}^+)^L \quad (7)$$

Trong công thức (7), $\alpha \in [0,1]$ là hằng số dùng để điều chỉnh mức độ ưu tiên cho mỗi loại đường đi. Nếu ưu tiên cho các đường đi có trọng số dương ta chọn α gần với *0*. Nếu ưu tiên cho các đường đi có trọng số âm ta chọn α gần với *1*. Nếu lấy $\alpha = 0$ mô hình trở lại đúng mô hình của Huang [4]. Thuật toán dự đoán trên mô hình đồ thị hai phía có trọng số được mô tả chi tiết trong Hình 3.

Bước 1 (*Khởi tạo*):
- Tạo lập ma trận trọng số $W = (w_{ix})$.
- Tạo lập ma trận (w_{ix}^+) biểu diễn G^+.
- Tạo lập ma trận (w_{ix}^+) biểu diễn G^-.

Bước 2 (*Tính toán trên đồ thị G^+*):

$$(w_{ix}^+)^L = \begin{cases} (w_{ix}^+) & \text{nếu } L=1 \\ (w_{ix}^+).(w_{ix}^+)^T.(w_{ix}^+)^{L-2} & \text{nếu } L=3,5,7,... \end{cases}$$

Bước 3 (*Tính toán trên đồ thị G^-*):

$$(w_{ix}^-)^L = \begin{cases} (w_{ix}^-) & \text{nếu } L=1 \\ (w_{ix}^-).(w_{ix}^-)^T.(w_{ix}^-)^{L-2} & \text{nếu } L=3,5,7,... \end{cases}$$

Bước 4 (*Kết hợp trọng số trên cả hai đồ thị*):

$$w_{ix}^L = \alpha.(w_{ix}^-)^L + (1-\alpha).(w_{ix}^+)^L$$

Bước 5 (*Sắp xếp theo thứ tự giảm dần của w^L_{ix}*).

Bước 6 (*Sinh ra tư vấn cho người dùng i*).
<Chọn K sản phẩm có $w_{ix}>0$ có giá trị lớn nhất tư vấn cho người dùng i>

Hình 3. Thuật toán dự đoán trên đồ thị hai phía

Độ phức tạp của thuật toán phụ thuộc vào L phép toán nhân ma trận cấp $N \times M$. Sử dụng thuật toán nhân hai ma trận hiệu quả nhất hiện nay của *Coppersmith–Winograd* sẽ cho ta độ phức tạp là $O(N^{2.376})$[4]. Để tránh các phép nhân ma trận có kích cỡ lớn, chúng tôi sử dụng thuật toán lan truyền mạng có độ phức tạp là $O(N.S)$, trong đó N là số lượng người dùng, S là số lượng trung bình các giá trị đánh giá khác ∅ của người dùng [1].

IV. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Để thấy rõ hiệu quả của mô hình đề xuất, chúng tôi thực hiện tiến hành thử nghiệm trên hai bộ dữ liệu MovieLens [11] và BookCrossing [12]. Trong đó, tập dữ liệu MovieLens được biểu diễn bằng 5 mức đánh giá, tập dữ liệu BookCrossing được biểu diễn bằng 10 mức đánh giá. Sai số dự đoán được ước lượng thông qua độ chính xác (*precision*), độ nhạy (*recall*) và tỉ lệ *F-Measure* theo thủ tục được mô tả dưới đây.

IV.1. Dữ liệu thử nghiệm

Tập dữ liệu MovieLens gồm 1682 người dùng, 942 phim với trên 100,000 đánh giá, các mức đánh giá được thiết lập từ 1 đến 5, mức độ thưa thớt dữ liệu đánh giá là 98.7%. Các mức đánh giá 4, 5 được chuyển đổi thành 0.1, 0.2. Các mức đánh giá 3, 2, 1 được dịch chuyển thành 0.0, -0.1, -0.2.

Tập dữ liệu BookCrossing là cơ sở dữ liệu bao gồm 278,858 người dùng với 1,031,175 đánh giá cho 271,065 đầu sách. Các mức đánh giá được thiết lập từ 0 đến 1.0, trung bình số lượng sách người dùng chưa đánh giá là 99.1%. Các mức đánh giá từ 0.6 đến 1.0 được dịch chuyển thành 0.1 đến 0.5 theo thứ tự. Các mức đánh giá từ 0.5 đến 0.0 được dịch chuyển thành 0.0, -0.1,…,-0.5 theo thứ tự.

Việc chuyển đổi dữ liệu theo ngưỡng θ=3 đối với tập dữ liệu MovieLans và θ=5 đối với bộ dữ liệu BookCrossing là cách làm phổ biến của các tác giả trước đây trong khi xem xét bài toán lọc cộng tác như bài toán phân loại hai lớp (-1,1)[1, 3, 4, 9]. Trong mô hình này, chúng tôi xem xét bài toán lọc cộng tác như bài toán phân loại nhiều lớp. Mỗi lớp thuộc một nhóm nhãn phân loại khác nhau trong khoảng [-1,1]. Chúng tôi cũng không chọn giá trị nhãn phân loại cực đại (1) hoặc cực tiểu (-1) vì phương pháp dự đoán chỉ quan tâm đến giá trị dự đoán lớn hay bé trong quá trình huấn luyện. Do vậy, sử dụng các giá trị nhãn phân loại nhỏ hơn 1 tiện lợi và chính xác hơn rất nhiều trong khi so sánh kết quả dự đoán.

VI.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.

Phương pháp ước lượng sai số dự đoán cho lọc cộng tác được sử dụng phổ biến là độ đo trung bình sai số tuyệt đối (MAE) [8]. Tuy nhiên, độ đo này chỉ được áp dụng với các phương pháp dự đoán có cùng miền xác định với giá trị đánh giá. Chính vì vậy, trong kiểm nghiệm này chúng tôi sử dụng phương pháp ước lượng sai số dự đoán thông qua độ chính xác *(precision)*, độ nhạy *(recall)* và *F-Measure* xác định theo công thức (8), (9), (10). Đây cũng là một phương pháp kiểm nghiệm được nhiều tác giả sử dụng cho lọc cộng tác [8].

$$P = \frac{N_{rs}}{N_r} \tag{8}$$

$$R = \frac{N_{rs}}{N} \tag{9}$$

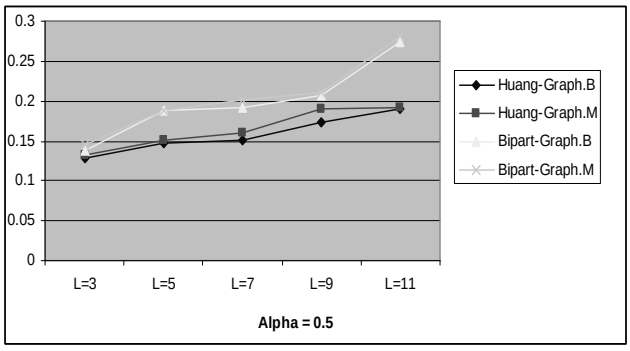
$$F - Measure = \frac{2 \times P \times R}{(P + R)} \tag{10}$$

Ở đây, N là tổng số các đánh giá người dùng trong tập dữ liệu kiểm tra trong đó có N_r là số các sản phẩm người dùng đã đánh giá thích hợp, N_{rs} là số các sản phẩm phương pháp lọc dự đoán chính xác. Giá trị *P, R, F_Measure* càng lớn độ chính xác của phương pháp càng cao.

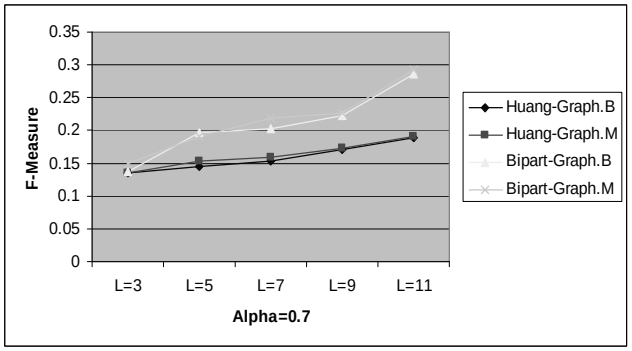
IV.3. Kết quả thử nghiệm

Để đánh giá hiệu quả của phương pháp đề xuất (ký hiệu là Bipart-Graph), chúng tôi tiến hành hai thử nghiệm trên các tập dữ liệu nêu trên. Thử nghiệm thứ nhất nhằm đánh giá ảnh hưởng của các đánh giá có trọng số âm và độ dài đường đi *L* đối với thói quen sử dụng sản phẩm của người dùng. Thử nghiệm này được so sánh với mô hình đồ thị hai phía của Huang (Ký hiệu là Huang-Graph[4]). Thử nghiệm thứ hai nhằm đánh giá kết quả dự đoán so với các phương pháp lọc khác, đặc biệt là kết quả dự đoán trong trường hợp dữ liệu thưa.

Đối với thử nghiệm thứ nhất, chúng tôi giữ lại tất cả các đánh giá có trọng số âm và trọng số dương trên cả hai tập dữ liệu. Chọn $\alpha = 0.5$, sau đó thực hiện quá trình huấn luyện nêu trên theo độ dài đường đi *L*. Kết quả được chỉ ra trên Hình 4, Bảng 5 cho thấy, khi L tăng (*L=3, 5, 7, 9, 11*) giá trị F-Measure của các mô hình đều tăng. Điều đó chứng tỏ việc suy diễn theo độ dài đường đi trên đồ thị cho phép ta tận dụng được các mối quan hệ gián tiếp giữa các người dùng khác nhau để tăng cường vào kết quả dự đoán.



Hình 4. Biến đổi của F-Measure với α=0.5



Hình 5. Biến đổi của F-Measure với α=0.7

Tiếp đến, chúng tôi chọn α=0.7 cho mô hình đồ thị đề xuất và thực hiện huấn luyện theo đường đi độ dài *L =3, 5, 7, 9, 11* (Hình 5, Bảng 6). Kết quả cho thấy, F-Measures của các mô hình đều tăng nhưng mô hình đề xuất cho lại kết quả tốt hơn rất nhiều so với mô hình của Huang [4]. Lý do khi α=0.7 kết quả dự đoán của phương pháp được cải thiện hơn vì số lượng các đánh giá dương lớn hơn rất nhiều lần số lượng các đánh giá âm trong các tập dữ liệu huấn luyện. Do vậy, với α =0.5 các đường đi có trọng số âm không ảnh hưởng nhiều đến các đường đi có trọng số dương. Điều đó chứng tỏ, đối với các đánh giá âm ta không được phép bỏ qua mà còn phải được chú ý đến nó nhiều hơn trong quá trình huấn luyện.

Bảng 5. Giá trị của F-Measure với α=0.5

Phương pháp	Độ dài đường đi				
	L=3	*L=5*	*L=7*	*L=9*	*L=11*
Huang-Graph.B	0.1279	0.1464	0.1511	0.1727	0.1899
Huang-Graph.M	0.1315	0.1513	0.1607	0.1893	0.1915
Bipart-Graph.B	0.1373	0.1877	0.1911	0.2073	0.2732
Bipart-Graph.M	0.1458	0.1889	0.2012	0.2102	0.2821

Bảng 6. Giá trị của F-Measure với α=0.7

Phương Pháp	Độ dài đường đi				
	L=3	*L=5*	*L=7*	*L=9*	*L=11*
Huang-Graph.B	0.1352	0.1457	0.1531	0.1718	0.1899
Huang-Graph.M	0.1356	0.1531	0.1598	0.1732	0.1905
Bipart-Graph.B	0.1378	0.1971	0.2031	0.2237	0.2873
Bipart-Graph.M	0.1485	0.1909	0.2188	0.2271	0.2914

Thử nghiệm thứ hai được thực hiện nhằm so sánh đánh giá kết quả với các phương pháp: Lọc cộng tác dựa vào người dùng (User Based) [9], lọc cộng tác dựa vào sản phẩm (Item Based) [2] và lọc cộng tác dựa vào mô hình đồ thị của Huang. Trong đó thử nghiệm này chúng tôi thực hiện với α =0.5, L=11. Độ chính xác, độ nhạy và F-Measure được lấy trung bình từ 10 lần kiểm nghiệm ngẫu nhiên dựa trên các tập dữ liệu kiểm tra dưới đây:

- Tập *Test1.M, Test1.B* (M ký hiệu cho tập tập MovieLans, B ký hiệu cho tập tập BookCrossing): Loại bỏ ngẫu nhiên các giá trị đánh giá trong mỗi tập dữ liệu tương ứng sao cho mỗi người dùng chỉ còn lại 5 đánh giá biết trước. Trường hợp này được xem là trường hợp dữ liệu rất thưa.
- Tập *Test2.M, Test2.B:* Loại bỏ ngẫu nhiên các giá trị đánh giá trong mỗi tập dữ liệu tương ứng sao cho mỗi người dùng chỉ còn lại 10 đánh giá biết trước. Trường hợp này cũng được xem là trường hợp dữ liệu rất thưa.
- Tập *Test3.M, Test3.B:* Loại bỏ ngẫu nhiên các giá trị đánh giá sao trong mỗi tập dữ liệu tương ứng cho mỗi người dùng chỉ còn lại 15 đánh giá biết trước. Trường hợp này được xem là trường hợp dữ liệu thưa.
- Tập *Test4.M Test4.B:* Loại bỏ ngẫu nhiên các giá trị đánh giá trong mỗi tập dữ liệu tương ứng sao cho mỗi người dùng chỉ còn lại ít nhất 20 đánh giá biết trước. Trường hợp này được xem là trường hợp có tương đối đầy đủ dữ liệu.

Bảng 7. *Kết quả kiểm nghiệm trên tập MovieLens*

Phương pháp	Độ đo	Số đánh giá biết trước trong tập kiểm tra			
		5	10	15	20
UserBased	Độ nhạy	0.144	0.157	0.162	0.279
	Độ chính xác	0.174	0.186	0.198	0.218
	F-Measure	0.158	0.170	0.178	0.245
ItemBased	Độ nhạy	0.098	0.118	0.144	0.259
	Độ chính xác	0.144	0.174	0.211	0.244
	F-Measure	0.117	0.141	0.171	0.251
Huang-Graph	Độ nhạy	0.142	0.165	0.234	0.381
	Độ chính xác	0.175	0.234	0.292	0.339
	F-Measure	**0.157**	**0.194**	**0.299**	**0.359**
Bipart-Graph	Độ nhạy	0.198	0.215	0.312	0.397
	Độ chính xác	0.211	0.284	0.325	0.377
	F-Measure	**0.204**	**0.245**	**0.318**	**0.387**

Kết quả kiểm nghiệm được trên các tập dữ liệu thể hiện trong Bảng 7, Bảng 8 cho thấy phương pháp đề xuất cho lại kết quả dự đoán tốt hơn rất nhiều so với các phương pháp khác. Điều đó có thể lý giải mô hình đề xuất đã tìm ra và tích hợp được ngữ nghĩa ẩn của các mối quan hệ gián tiếp giữa người dùng và sản phẩm để tăng cường thêm vào kết quả dự đoán. Một lợi thế khác cũng cần được nhắc đến là phương pháp tiếp cận của mô hình khá đơn giản và dễ cài đặt cho các hệ thống lọc cộng tác.

Bảng 8. *Kết quả kiểm nghiệm trên tập BookCrossing*

Phương pháp	Độ đo	Số đánh giá biết trước trong tập kiểm tra			
		5	10	15	20
UserBased	Độ nhạy	0.102	0.121	0.142	0.149
	Độ chính xác	0.174	0.194	0.214	0.265
	F-Measure	0.129	0.149	0.171	0.191
ItemBased	Độ nhạy	0.092	0.114	0.124	0.152
	Độ chính xác	0.147	0.163	0.211	0.259
	F-Measure	0.113	0.134	0.156	0.192
Huang-Graph	Độ nhạy	0.113	0.129	0.134	0.156
	Độ chính xác	0.248	0.286	0.310	0.326
	F-Measure	**0.155**	**0.178**	**0.187**	**0.211**
Bipart-Graph	Độ nhạy	0.125	0.138	0.157	0.185
	Độ chính xác	0.287	0.256	0.234	0.473
	F-Measure	**0.174**	**0.179**	**0.188**	**0.266**

V. KẾT LUẬN

Kết quả kiểm nghiệm trên các bộ dữ liệu thực về sách và phim có nhiều mức đánh giá khác nhau cho thấy mô hình đề xuất cho lại độ chính xác, độ nhạy và tỷ lệ F-Measure cao hơn hẳn các phương pháp ItemBased, UserBased và Huang-Graph. Điều đó có thể khẳng định, phương pháp biểu diễn và dự đoán của mô hình đồ thị hai phía có trọng số đề xuất cải thiện đáng kể chất lượng dự đoán cho lọc cộng tác. Ưu điểm nổi bật của mô hình so với những mô hình trước đây là thỏa mãn biểu diễn hiện có của tất cả các tập dữ liệu hiện nay của lọc cộng tác.

Phương pháp dự đoán được đưa về bài toán tìm kiếm trên đồ thị có trọng số cho phép ta phân biệt được mức độ quan trọng của từng loại đường đi bằng

cách sử dụng các thuật toán hiệu quả đã được áp dụng thành công cho nhiều ứng dụng khác nhau trên đồ thị. Chất lượng dự đoán được cải thiện bằng cách mở rộng các đường đi từ đỉnh người dùng đến đỉnh sản phẩm. Điều này cho phép ta tận dụng được các mối liên hệ gián tiếp giữa người dùng và sản phẩm vào quá trình dự đoán.

TÀI LIỆU THAM KHẢO

[1] NGUYEN DUY PHUONG, LE QUANG THANG, TU MINH PHUONG (2008), *"A Graph-Based for Combining Collaborative and Content-Based Filtering"*. PRICAI 2008: 859-869.

[2] X. SU, T. M. KHOSHGOFTAAR (2009), *"A Survey of Collaborative Filtering Techniques"*. Advances in Artificial Intelligence, vol 2009, pp.1-20.

[3] Z. HUANG, D. ZENG, H. CHEN (2007), *"Analyzing Consumer-product Graphs: Empirical Findings and Applications in Recommender Systems"*, Management Science, 53(7), 1146-1164.

[4] Z. HUANG, H. CHEN, D. ZENG (2004), *"Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering"*, ACM Transactions on Information Systems, vol. 22(1) pp. 116–142

[5] T. HOFMANN (2004), *"Latent Semantic Models for Collaborative Filtering"*, ACM Trans. Information Systems, vol. 22, No. 1, pp. 89-115.

[6] C.C.AGGARWAL, J.L. WOLF, K.L. WU, AND P.S.YU (1999), *"Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering"*, Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.

[7] R. JIN, L. SI, AND C. ZHAI (2003), *"Preference-Based Graphic Models for Collaborative Filtering"*, Proc. 19th Conf. Uncertainty in Artificial Intelligence (UAI 2003).

[8] J.L. HERLOCKER, J.A. KONSTAN, L.G. TERVEEN, AND J.T. RIEDL (2004), *"Evaluating Collaborative Filtering Recommender Systems"*, ACM Trans. Information Systems, vol. 22, No. 1, pp. 5-53.

[9] J. S. BREESE, D. HECKERMAN, AND C. KADIE (1998), *"Empirical analysis of Predictive Algorithms for Collaborative Filtering"*, In Proc. of 14th Conf. on Uncertainty in Artificial Intelligence, pp. 43-52.

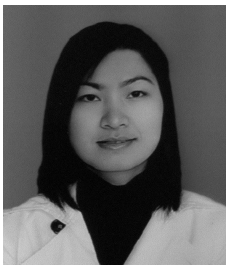
[10] G. ADOMAVICIUS, A. TUZHILIN (2005), *"Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions"*, IEEE Transactions On Knowledge And Data Engineering, vol. 17, No. 6, 2005.

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

[12] http://www.grouplens.org/node/74

Nhận bài ngày: 11/04/2012

SƠ LƯỢC TÁC GIẢ



MAI THỊ NHƯ

Sinh ngày 06/08/1984 tại Hà Nội.

Tốt nghiệp đại học và cao học tại Học viện Công nghệ Bưu chính Viễn thông vào năm 2007 và 2012.

Hiện đang công tác tại đang công tác tại Công ty máy tính HP Việt Nam.

Hướng nghiên cứu: học máy ứng dụng trong lọc thông tin. Điện thoại : 0904941166,

Email: mtnhu@yahoo.com

NGUYỄN DUY PHƯƠNG

Sinh ngày 20/02/1965 tại Hà Nội.

Tốt nghiệp đại học và cao học tại Đại học Tổng hợp Hà Nội vào năm 1988 và 1997. Bảo vệ luận án tiến sỹ tại Đại học Quốc Gia Hà Nội năm 2011.

Hiện đang công tác tại Học viện Công nghệ Bưu chính Viễn thông.

Hướng nghiên cứu: học máy ứng dụng trong lọc thông tin. Điện thoại : 0913575442

Email: phuongnd@ptit.edu.vn