Lọc cộng tác với độ đo tương tự dựa trên đồ thị

Collaborative Filtering with a Graph-based Similarity Measure

Nguyễn Duy Phương và Từ Minh Phương

Abstract: *Collaborative filtering is a technique widely used in recommender systems. Based on the behaviors of users with similar taste, the technique can predict and recommend products the current user is likely interested in, thus alleviates the information overload problem for Internet users. The most popular collaborative filtering approach is based on the similarity between users, or between products. The quality of similarity measure, therefore, has a large impact on the recommendation accuracy. In this paper, we propose a new similarity measure based on graph models. The similarity between two users (or symmetrically, two products) is computed from connections on a graph with vertices beeing users and products. The computed similarity measure is then used with k – nearest neighbor algorithm to generate predictions. Empirical results on real movie datasets show that the proposed method significantly outperforms both collaborative filtering with traditional similarity measures and pure graph-based collaborative filtering.*

I. MỞ ĐẦU

Khó khăn lớn với người sử dụng Internet và các dịch vụ thương mại điện tử là luôn có quá nhiều phương án để lựa chọn. 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 lựa chọn* (recommender systems) cho phép phần nào giải quyết vấn đề này bằng cách dự đoán và cung cấp cho người dùng một danh sách ngắn các sản phẩm, bản tin, phim, video, v.v… mà nhiều khả năng người dùng sẽ quan tâm. Hiện nhiều hệ tư vấn thương mại đã được sử dụng rất thành công như hệ thống của Amazon, Netflix, Yahoo!, Youtube.

Có hai kỹ thuật chính được sử dụng trong tư vấn lựa chọn: *lọc theo nội dung* (content-based filtering) và *lọc cộng tác* (collaborative filtering) [2]. Lọc theo nội dung phân tích đặc trưng nội dung các sản phẩm mà người dùng đã chọn trong quá khứ và tư vấn cho người dùng những sản phẩm mới có đặc trưng nội dung tương tự. Để sử dụng được phương pháp này, nội dung sản phẩm phải được mô tả rõ ràng dưới dạng văn bản hoặc thông qua một số đặc trưng. Trái lại, lọc cộng tác dựa trên nhóm người dùng đã từng chọn những sản phẩm giống người dùng cần tư vấn để xác định sản phẩm cần giới thiệu với người này. So với lọc theo nội dung, lọc cộng tác có ưu điểm là không đòi hỏi sản phẩm phải được mô tả dưới dạng văn bản hay đặc trưng. Kết quả thử nghiệm cũng cho thấy, lọc cộng tác lọc tốt hơn lọc nội dung trong nhiều trường hợp [2]. Trong bài báo này, chúng tôi tập trung vào phương pháp lọc cộng tác.

Phương pháp lọc cộng tác điển hình được áp dụng rộng rãi nhất là phương pháp *k* – láng giềng gần nhất. Phương pháp này còn được gọi là *lọc dựa trên bộ nhớ* (memory-based filtering) [3,4,6,7] để phân biệt với *lọc dựa trên mô hình* (model-based filtering) [8,11,12]. Với mỗi người dùng, hệ thống xác định *k* người dùng có sở thích giống người đó nhất dựa trên những sản phẩm họ đã chọn hoặc đã đánh giá trong quá khứ, sau đó tư vấn cho người dùng hiện thời sản phẩm mà k người này đã chọn. Tương tự như vậy, thay vì tìm *k* người dùng gần nhất, ta cũng có thể tìm k láng giềng gần nhất cho mỗi sản phẩm và dựa trên việc người dùng có quan tâm tới các láng giềng này trong quá khứ không để quyết định lựa chọn hoặc không lựa chọn sản phẩm đang xét. Trong trường hợp thứ nhất, lọc cộng tác được gọi là *lọc dựa trên người dùng* (user-based collaborative filtering), trong trường hợp

thứ hai là lọc *dựa trên sản phẩm* (item-based).

Để lọc cộng tác dựa trên bộ nhớ cho kết quả tốt cần xác định chính xác độ tương tự giữa người dùng từ ma trận đánh giá của người dùng đối với sản phẩm (hoặc độ tương tự giữa các sản phẩm, tùy theo phương pháp nào được sử dụng). Thông thường, độ đo tương tự được sử dụng là độ đo tương tự giữa hai vectơ như cosin hay độ tương quan Pearson. Tuy nhiên, các độ đo này cho kết quả không tốt trong trường hợp dữ liệu thưa thớt, tức là khi mỗi người dùng chỉ lựa chọn hoặc đánh giá ít sản phẩm trong quá khứ - tình huống điển hình đối với các hệ thống sử dụng lọc cộng tác.

Để giảm bớt ảnh hưởng của vấn đề dữ liệu thưa tới hiệu quả lọc cộng tác dựa trên bộ nhớ, nhiều phương pháp đã được đề xuất như kỹ thuật làm trơn nhờ phân cụm [14], kết hợp lọc dựa trên người dùng với dựa trên sản phẩm [15], và đặc biệt là dựa trên quan hệ kết hợp từ đồ thị người dùng – sản phẩm [9,10].

Trong bài báo này, chúng tôi đề xuất một phương pháp mới tính toán mức độ tương tự giữa các cặp người dùng hoặc sản phẩm có độ ổn định tốt hơn khi độ thưa thớt dữ liệu thay đổi. Dựa trên đồ thị người dùng – sản phẩm, phương pháp đề xuất xác định độ liên thông dưới dạng các đường đi có trọng số giữa các cặp người dùng (hoặc sản phẩm). Độ liên thông này sau đó được sử dụng như độ tương tự khi xác định k láng giềng gần nhất. Thuật toán đề xuất cho phép tính toán độ dài nhỏ nhất của đường đi đủ đảm bảo có độ phủ tốt trong trường hợp dữ liệu thưa. Phương pháp đề xuất tương tự phương pháp của Huang et al. [10] ở chỗ đều dựa trên đồ thị người dùng – sản phẩm. Tuy nhiên, khác với phương pháp trong [10], chúng tôi không sử dụng trực tiếp mức độ liên kết giữa người dùng với sản phẩm để đưa ra dự đoán. Thay vào đó, liên kết người dùng với người dùng hoặc sản phẩm với sản phẩm được sử dụng tính độ tương tự và dùng với mô hình dựa trên bộ nhớ. Việc kết hợp hai phương pháp đồ thị và k láng giềng tạo hiệu ứng làm trơn và cho kết quả thực nghiệm tốt hơn đáng kể so với từng phương pháp riêng rẽ. Ngoài ra, so với phương pháp của Huang và cộng sự, phương pháp đề xuất có bước xác định rõ ràng độ dài cần thiết của đường đi để đảm bảo độ phủ tốt khi dữ liệu thưa.

Phương pháp đề xuất trong bài báo được thử nghiệm trên các bộ dữ liệu thực tế về đánh giá của người dùng đối với phim. Kết quả thử nghiệm cho thấy việc phương pháp cho kết quả lọc tốt hơn so với phương pháp lọc cộng tác dựa trên các độ đo tương quan hiện nay, cũng như phương pháp đồ thị thuần túy [10].

II. BÀI TOÁN LỌC CỘNG TÁC

Bài toán lọc cộng tác có thể phát biểu như sau. Cho tập hợp U gồm N người dùng $U = \{u_1, u_2,\ldots, u_N\}$, và là tập P gồm M sản phẩm $P = \{p_1, p_2,.., p_M\}$. Mỗi sản phẩm $p_x \in P$ có thể là bài báo, bản tin, hàng hóa, phim, ảnh, dịch vụ, v.v… 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 $r_{ix} \in \{\varnothing, 1, 2, ..,G\}$ là đánh giá của người dùng $u_i \in U$ đối với sản phẩm $p_x \in P$. 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. Chẳng hạn nếu trong quá khứ người dùng đã từng mua sản phẩm hoặc xem trang web thì đánh giá của người dùng với sản phẩm đó sẽ có giá trị là 1. Giá trị $r_{ix} = \varnothing$ trong trường hợp người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_x. Nhiệm vụ của lọc cộng tác là dự đoán đánh giá của người dùng hiện thời $u_a \in U$ đối với những mặt hàng mới $p_x \in P$, trên cơ sở đó tư vấn cho người dùng u_a những sản phẩm được đánh giá cao [1].

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

	p_1	p_2	p_3	p_4
u_1	5	∅	4	∅
u_2	∅	3	4	∅
u_3	∅	3	∅	2

Bảng 1 là một ví dụ về ma trận đánh giá cho hệ lọc cộng tác gồm *3* người dùng $U = \{ u_1, u_2, u_3\}$ và 4 sản phẩm $P = \{p_1, p_2, p_3, p_4\}$. Các giá trị đánh giá được biểu diễn có giá trị $r_{ix} \in \{\varnothing, 1, 2, 3, 4, 5\}$. Những giá trị $r_{ix}=\varnothing$ được hiểu là người dùng $i \in U$ chưa biết đến

sản phẩm $p_x \in P$. Để tư vấn, chẳng hạn cho người dùng u_3, thuật toán lọc cộng tác phải xác định giá trị cho các ô trống trong dòng tương ứng với u_3.

II.1. Lọc cộng tác dựa trên bộ nhớ

Có nhiều phương pháp khác nhau được đề xuất và sử dụng trên thực tế cho bài toán lọc cộng tác. Su và Khoshgoftaar [1] phân loại các phương pháp giải quyết bài toán lọc cộng tác thành hai cách tiếp cận chính: Lọc cộng tác dựa vào bộ nhớ (Memory-Based [3, 4, 6]) và Lọc cộng tác dựa vào mô hình (Model-Based [8, 11, 12]). Lọc dựa vào bộ nhớ được thực hiện theo hai phương pháp chính: lọc dựa vào người dùng (UserBased) và lọc dựa vào sản phẩm (ItemBased) [1, 2]. Đặc điểm chung của cả hai phương pháp này đều dựa vào các độ đo khoảng cách (Euclid, Minkowski...), độ đo tương tự (Cosin, Entropy,...), độ đo tương quan (Pearson, Root Mean Square, Spearman Rank, Kendall,...) tính toán mức độ tương tự giữa các cặp người dùng (hoặc sản phẩm) để tìm ra các sản phẩm có mức độ tương tự cao phù hợp cho mỗi người dùng [7, 16].

Về bản chất, lọc cộng tác dựa trên bộ nhớ tương tự phương pháp *k* láng giềng gần nhất trong học máy. Trong trường hợp lọc theo người dùng (UserBased), phương pháp này được thực hiện qua các bước sau:

1) Tính toán mức độ tương tự giữa các cặp người dùng. Các độ đo được sử dụng rộng rãi nhất để xác định độ tương tự giữa hai người dùng hoặc hai sản phẩm là độ tương quan Pearson và cosin giữa hai vectơ.
2) Xác định tập *k* láng giềng cho người dùng hiện thời.
3) Tổ hợp các đánh giá của *k* láng giềng gần nhất đối với sản phẩm mà người dùng hiện thời chưa biết để dự đoán đánh giá của người dùng hiện thời cho sản phẩm này. Cách tổ hợp đơn giản nhất là lấy trung bình cộng theo *k* láng giềng, hoặc có thể tổ hợp theo nhiều dạng trọng số khác nhau.
4) Trả về cho người dùng hiện thời các sản phẩm có đánh giá cao nhất.

Trong trường hợp lọc theo sản phẩm (ItemBased), thay vì tìm *k* láng giềng cho người dùng hiện thời, hệ thống sẽ tìm *k* láng giềng gần nhất cho sản phẩm cần dự đoán, sau đó tổ hợp các đánh giá đã có của người dùng hiện thời đối với các láng giềng này để xác định đánh giá của người dùng hiện thời đối với sản phẩm cần dự đoán.

Mặc dù lọc cộng tác dựa trên bộ nhớ đơn giản và hiệu quả, việc áp dụng trên thực tế gặp khó khăn do vấn đề thưa thớt dữ liệu. Đối với các hệ thống lọc cộng tác, mỗi người dùng thường chỉ đánh giá rất ít sản phẩm do vậy đa số phần tử của ma trận đánh giá có giá trị rỗng. Khi thực hiện tính toán mức độ tương tự giữa các cặp người dùng u_{ij}, các độ đo tương quan chỉ thực hiện tính toán trên tập $P_{ij} \neq \varnothing$. Những sản phẩm có giá trị đánh giá khác $\varnothing$ ngoài tập P_{ij} sẽ không được tham gia vào quá trình tính toán. Điều này làm cho nhiều người dùng có sở thích tương tự nhau nhưng lại không xác định được bằng các độ đo tương quan do chưa cùng đánh giá một số sản phẩm. Ngược lại, nhiều cặp người dùng kém tương tự nhau nhưng vẫn được xác định trong tập láng giềng.

II.2. Lọc cộng tác sử dụng mô hình đồ thị

Để giảm ảnh hưởng của vấn đề dữ liệu thưa đối với lọc cộng tác dựa trên bộ nhớ, một số giải pháp đã được đề xuất, trong đó đáng chú ý là giải pháp sử dụng tính liên thông và bắc cầu trên đồ thị do Huang và cộng sự [10] đề xuất (để tiện cho việc trình bầy, phương pháp này sẽ được gọi là GraphBased trong phần còn lại của bài báo). Theo phương pháp này, ma trận người dùng – sản phẩm được sử dụng để xây dựng đồ thị với các đỉnh là người dùng và sản phẩm. Một đỉnh người dùng được nối với một đỉnh sản phẩm nếu người dùng đã mua hoặc đánh giá tốt về sản phẩm đó. Lưu ý là phương pháp này được đề xuất cho trường hợp ma trận đánh giá có 2 giá trị: 1 nếu người dùng đã chọn sản phẩm, và $\varnothing$ trong trường hợp ngược lại. Đồ thị trong phương pháp do Huang và cộng sự đề xuất có dạng tương tự như trên Hình 1, tuy nhiên tất cả các cạnh có trọng số bằng 1.

Các tác giả của GraphBased xác định mức độ quan tâm của người dùng hiện thời với một sản phẩm bằng cách tính tổng trọng số các đường đi có độ dài không lớn hơn *L* giữa người dùng và sản phẩm đó. Ở đây *L* là tham số của phương pháp và có giá trị lẻ do chỉ xét các đường đi bắt đầu từ nút người dùng và kết thúc ở nút sản phẩm. Trọng số đường đi độ dài *l* được tính bằng α^l, trong đó $0 < \alpha < 1$ là tham số của phương pháp. Do $\alpha < 1$ nên đường đi càng dài có trọng số càng nhỏ và do vậy đóng góp càng ít vào liên kết giữa người dùng với sản phẩm. Các sản phẩm có liên kết lớn nhất với người dùng hiện thời với độ liên kết xác định như trên sẽ được tư vấn cho người dùng hiện thời. Huang và cộng sự đã thử nghiệm phương pháp với dữ liệu thực và cho kết quả tốt hơn nhiều so với lọc cộng tác dựa trên bộ nhớ truyền thống, đặc biệt khi dữ liệu bị thưa thớt.

III. LỌC CỘNG TÁC SỬ DỤNG ĐỘ TƯƠNG TỰ DỰA TRÊN ĐỒ THỊ

Như đã trình bầy ở trên, việc tính toán độ tương tự từ ma trận đánh giá có thể gặp khó khăn khi dữ liệu thưa. Phương pháp đồ thị mô tả trong [9,10] cho phép giải quyết phần nào vấn đề đó bằng cách tận dụng các liên kết gián tiếp giữa người dùng và sản phẩm. Tuy nhiên phương pháp này chỉ sử dụng cho trường hợp đánh giá có giá trị nhị phân (chẳng hạn "thích", "không thích"). Việc xác định độ dài *L* tối đa của liên kết tương đối khó khăn mặc dù đây là tham số ảnh hưởng nhiều tới hiệu quả dự đoán. Trong phần này, chúng tôi trình bầy phương pháp đề xuất, theo đó liên kết trên đồ thị được sử dụng để tính độ tương tự thay vì tính trực tiếp liên kết giữa người dùng với sản phẩm. Bên cạnh đó, phương pháp đề xuất cũng tổng quát hơn, có thể sử dụng cho trường hợp đánh giá nhận bất cứ giá trị nào. Chúng tôi cũng trình bày cách xác định độ dài tối đa của liên kết để có độ phủ tốt với dữ liệu cụ thể.

Các nội dung cụ thể trong phần này bao gồm: 1) Phương pháp biểu diễn đồ thị cho lọc cộng tác; 2) Phương pháp tính toán mức độ tương tự giữa các cặp người dùng (hoặc sản phẩm); và 3) Xây dựng thuật toán dự đoán đánh giá của người dùng đối với các sản phẩm dựa vào đồ thị.

III.1. Phương pháp biểu diễn đồ thị cho lọc cộng tác

Để thuận lợi cho việc xây dựng đồ thị, ta giả sử r_{ix} có thể nhận giá trị trong khoảng [0,1] hoặc giá trị rỗng, tức là.

$$r_{ix} = \begin{cases} \textit{Nếu người dùng i đánh giá } x \\ \textit{ở mức độ } v\ (0 \le v \le 1). \\ \textit{Nếu người dùng i chưa biết đến} \\ \textit{sản phẩm } x. \end{cases} \quad (1)$$

Đối với các bộ dữ liệu có giá trị đánh giá $r_{ix} \in \{1, 2, .., g\}$ như phát biểu trong phần 2, ta chỉ cần thực hiện phép biến đổi đơn giản chuyển $r_{ix} = \frac{r_{ix}}{g}$. Phép biến đổi này vẫn bảo toàn được mức độ đánh giá theo thứ tự khác nhau của các hệ lọc cộng tác đồng thời tổng quát hơn cách biểu diễn nhị phân. Ví dụ, theo cách biểu diễn này, dữ liệu đánh giá trong Bảng 1 sẽ được chuyển thành Bảng 2.

Bảng 2. Ma trận đánh giá được chuyển đổi.

	p_1	p_2	p_3	p_4
u_1	1.0	∅	0.8	∅
u_2	∅	0.6	0.8	∅
u_3	∅	0.6	∅	0.4

Hệ lọc cộng tác với ma trận đánh giá $\{r_{ix}\}$ có thể biểu diễn dưới dạng đồ thị hai phía $G = <V, E>$, một phía là tập người dùng, phía còn lại là tập sản phẩm. Tập đỉnh *V* của đồ thị được chia thành hai tập: tập người dùng và tập sản phẩm ($V = U \cup P$). Tập cạnh *E* của đồ thị được xác định theo công thức (2). Mỗi cạnh $e \in E$ có dạng $e = (i, x)$, trong đó $i \in U$ và $x \in P$. Không tồn tại các cạnh của nối giữa hai đỉnh người dùng hoặc cạnh nối giữa hai đỉnh sản phẩm. Trọng số của mỗi cạnh được xác định theo (3).

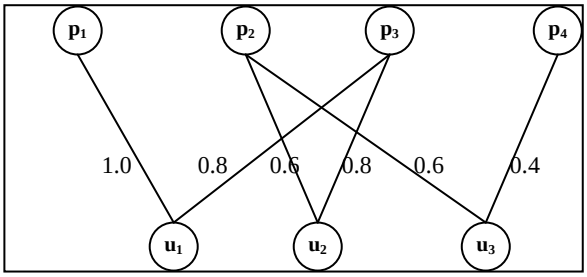
$$E = \{e = (i,x) : i \in U \wedge x \in P \mid r_{ix} \neq \phi\} \quad (2)$$

$$w_{ix} = \begin{cases} r_{ix} & if \quad (i,x) \in E \\ 0 & otherwise \end{cases} \quad (3)$$

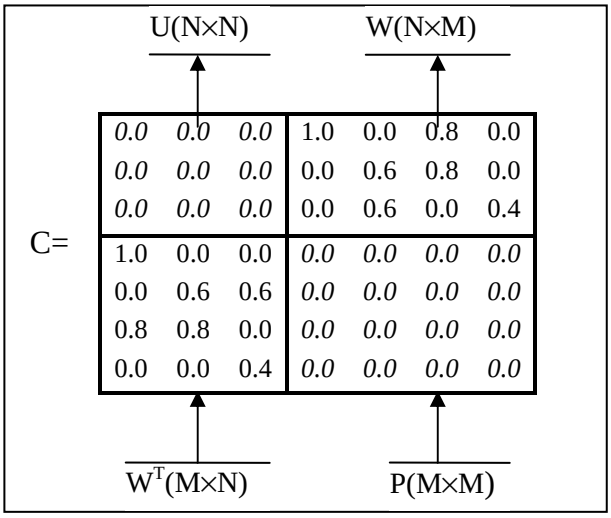
Gọi $C=\{c_{ij}\}$ là ma trận trọng số biểu diễn đồ thị G ($i = 1, 2,.., N+M; j = 1, 2, .., N+M$). Khi đó, ma trận vuông C sẽ được chia thành bốn phần theo công thức (4). Trong đó, ma trận vuông $U(N\times N)$ biểu diễn mối quan hệ giữa người dùng và người dùng, $P(M\times M)$ biểu diễn mối quan hệ giữa sản phẩm với sản phẩm, $W(N\times M)$ được xác định theo (7) biểu diễn mối quan hệ giữa người dùng và sản phẩm, $W^T(M\times N)$ là chuyển vị của $W(N\times M)$ biểu diễn mối quan hệ giữa sản phẩm và người dùng. Các phần tử của ma trận $U(N\times N)$, $P(M\times M)$ ban đầu đều có giá trị 0.

$$C=\begin{pmatrix} U(N\times N) & W(N\times M) \\ W^T(M\times N) & P(M\times M) \end{pmatrix} \quad (4)$$

Ví dụ, với hệ lọc cộng tác được cho trong Bảng 2, khi đó đồ thị hai phía biểu diễn cho lọc cộng tác được thể hiện trong Hình 1, thành phần của ma trận trọng số được thể hiện trong Hình 2.



Hình 1. Đồ thị hai phía biểu diễn cho lọc cộng tác.



Hình 2. Ma trận trọng số biểu diễn đồ thị hai phía.

Với đồ thị $G =<V,E>$ biểu diễn dưới dạng ma trận trọng số C xác định theo (4), lọc cộng tác có thể được xem xét như bài toán tìm kiếm trên đồ thị. Trong đó, mức độ tương tự giữa các cặp người dùng được tính toán dựa vào trọng số các đường đi từ đỉnh người dùng đến đỉnh người dùng, mức độ tương tự giữa các cặp sản phẩm được tính toán dựa vào trọng số các đường đi từ đỉnh sản phẩm đến đến đỉnh sản phẩm, mức độ phù hợp của người dùng đối với sản phẩm được tính toán dựa vào trọng số các đường đi từ đỉnh người dùng đến đỉnh sản phẩm. Mức độ tương tự giữa các cặp người dùng sẽ được điền giá trị vào ma trận $U(N\times N)$, mức độ tương tự giữa các cặp sản phẩm sẽ được điền giá trị vào ma trận $P(M\times M)$, mức độ phù hợp của người dùng đối với sản phẩm được điền giá trị vào ma trận $W(N\times M)$ và $W^T(M\times N)$. Trong đó, $U(N\times N)$, $P(M\times M)$, $W(N\times M)$ và $W^T(M\times N)$ được xác định theo (4). Nội dung cụ thể mỗi phương pháp tiếp cận sẽ được trình bày trong các mục tiếp theo của bài báo.

III.2. Lọc cộng tác sử dụng độ tương tự giữa cặp người dùng trên đồ thị

Độ đo tương tự cho người dùng dựa trên đồ thị

Như đã đề cập trong Mục II, các độ đo tương quan tính toán mức độ tương tự giữa các cặp người dùng $i\in U$, $j\in U$ dựa trên tập P_{ij} là tập các sản phẩm cả hai người dùng cùng đánh giá. Việc làm này dễ dàng được thực hiện trên đồ thị biểu diễn cho lọc cộng tác $G =<V, E>$ bằng cách tính toán tổng trọng số của tất cả các đường đi có độ dài *2* từ đỉnh $i\in U$ đến đỉnh $j\in U$. Trọng số của mỗi đường đi độ dài *2* từ đỉnh $i\in U$ đến đỉnh $j\in U$ được tính bằng tích trọng số của các cạnh tương ứng. Ví dụ để tính toán mức độ tương tự giữa người dùng u_1 và u_2 Hình 1, ta tính tổng trọng số các đường đi độ dài 2 từ đỉnh u_1 đến đỉnh u_2. Giữa u_1 và u_2 chỉ có duy nhất một đường đi độ dài 2: u_1-p_3-u_2. Trọng số của đường đi này được tính là 0.8*0.8 = 0.64. Tương tự như vậy, mức độ tương tự giữa người dùng u_2 và u_3 là 0.6*0.6 = 0.36. Như vậy, phương pháp tính toán mức độ tương tự giữa các cặp người dùng dựa trên các độ đo tương quan được xem như việc tính

toán tổng trọng số các đường đi độ dài *2* từ đỉnh người dùng đến đỉnh sản phẩm trên đồ thị $G = <V, E>$.

Khi tập các sản phẩm cả hai người dùng cùng đánh giá $P_{ij} = \emptyset$, các độ đo tương quan sẽ không xác định được mức độ tương tự giữa người dùng *i* và người dùng *j* [1, 7]. Ví dụ với người dùng u_1, u_3 trên Hình 1 sẽ không xác định được giá trị bằng các độ đo tương quan. Nguyên nhân của vấn đề này là do ma trận đánh giá quá thưa với phần lớn các phần tử có giá trị rỗng. Các giá trị đánh giá $r_{ij} = \emptyset$ chiếm 98.7% trên bộ dữ liệu MovieLens và 99.1% trên bộ dữ liệu BookCrossing. Tuy nhiên, quan sát những mối quan hệ khác giữa các cặp người dùng trên đồ thị ta có thể thấy họ vẫn tồn tại một mức độ tương tự tiềm ẩn nào đó. Ví dụ, giữa u_1 và u_3 đều tương tự với u_2 (Hình 1). Khai thác được những mối quan hệ này sẽ cải thiện đáng kể chất lượng dự đoán của các phương pháp UserBased và hạn chế được vấn đề dữ liệu thưa của lọc cộng tác.

Để khai thác được những mối quan hệ gián tiếp trên đồ thị, chúng tôi thực hiện mở rộng độ dài đường đi từ đỉnh người dùng đến đỉnh người dùng trên đồ thị. Ví dụ giữa u_1 và u_3 tồn tại đường đi độ dài *4*: u_1-p_3-u_2-p_2-u_3, trọng số của đường đi này là 0.8*0.8*0.6*0.6 = 0.2304. Vì *G* là đồ thị hai phía nên đường đi từ đỉnh người dùng đến đỉnh người dùng luôn là một số chẵn (2, 4, 6, 8,...). Mặt khác, trọng số mỗi cạnh của đồ thị là một số dương nhỏ hơn *1* nên các đường đi có độ dài lớn sẽ được đánh trọng số thấp, đường đi có độ dài nhỏ sẽ được đánh trọng số cao. Mức độ tương tự giữa người dùng $i \in U$ và người dùng $j \in U$ được ước lượng bằng tổng các trọng số của tất cả các đường đi độ dài *L* đi từ đỉnh *i* đến đỉnh *j* trên đồ thị. Bằng cách tiếp cận này mức độ tương tự giữa các cặp người dùng được xác định dựa trên tất cả các mối quan hệ trực tiếp hoặc gián tiếp.

Một cách tổng quát, gọi $U^L(N \times N)$ là tổng trọng số các đường đi có độ dài *L* từ đỉnh $i \in U$ đến đỉnh $j \in U$ trên đồ thị *G* ($i=1, 2,.., N; j=1, 2, .., N$). Vì tổng trọng số mỗi đường đi độ dài *L* được tính bằng tích của trọng số các cạnh, nên tổng trọng số các đường đi độ dài *L* từ đỉnh người dùng đến đỉnh người dùng trên đồ thị *G* được xác định theo công thức (5).

$$U^L = \begin{cases} W.W^T & if \quad L = 2 \\ W\,W^T\,U^{L-2} & if \quad L = 4,6,8,.. \end{cases} \quad (5)$$

Mức độ tương tự giữa các cặp người dùng được xác định theo (5) phụ thuộc vào độ dài đường đi L từ đỉnh người dùng đến đỉnh người dùng trên đồ thị. Do vậy, một vấn đề đặt ra là với mỗi người dùng $i \in U$ giá trị của *L* được lấy bằng bao nhiêu cho tốt. Ngược lại đồ thị $G = <V,E>$ phải có tính chất gì để ta có thể xác định được *L* sao cho mức độ tương tự giữa các cặp người dùng $u_{ij} \neq 0$. Định lý 1 dưới đây sẽ cho ta một cách xác định *L* trong trường hợp đồ thị biểu diễn của lọc cộng tác $G = <V,E>$ liên thông. Đối với các hệ lọc cộng tác có biểu diễn đồ thị $G = <V, E>$ không liên thông chúng tôi sẽ trình bày trong những kết quả nghiên cứu tiếp theo của bài báo.

Định lý 1. Nếu đồ thị biểu diễn cho các hệ lọc cộng tác $G = <V, E>$ liên thông thì luôn luôn tồn tại số tự nhiên chẵn *L* để $u_{ij}^L \neq 0$ với mọi $i, j \in U$. Trong đó, u_{ij}^L được xác định theo (5).

Chứng minh. Giả sử đồ thị biểu diễn cho các hệ lọc cộng tác $G = <V,E>$ liên thông. Khi đó luôn luôn tồn tại đường đi từ đỉnh $i \in U$ đến mọi $j \in U$ trên đồ thị. Vì $G = <V,E>$ là đồ thị hai phía được biểu diễn theo (4), nên luôn tồn tại số chẵn *L* sao cho từ $i \in U$ đến $j \in U$ được nối bằng đúng *L* cạnh. Do u_{ij}^L được xác định theo (5) là tổng trọng số các đường đi có độ dài *L*; Trọng số mỗi đường đi có độ dài *L* là tích của trọng số các cạnh có $w_{ij}^L \neq 0$, vì vậy nên $u_{ij}^L \neq 0$ là điều cần chứng minh.

Như vậy, để xác định mức độ tương tự giữa người dùng $i \in U$ với người dùng $j \in \{U \setminus i\}$, ta chỉ cần chọn giá trị *L nhỏ nhất* để $u_{ij}^L \neq 0$ với mọi $j \in U$.

Thuật toán lọc cộng tác

Dựa trên Định lý 1, chúng tôi đề xuất thuật toán UserBased-Graph cho lọc cộng tác như trình bầy trong Hình 3.

Đầu vào:

- Ma trận trọng số C là biểu diễn đồ thị $G = <V, E>$ cho lọc cộng tác .
- $i \in U$ là người dùng cần được tư vấn.
- K là số lượng người dùng của tập láng giềng.

Đầu ra:

- Dự đoán x: $r_{ix} \mid x \in P \backslash P_i$.(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:

$L \leftarrow 2$;//Khởi tạo độ dài đường đi ban đầu

Repeat

$$U^L = \begin{cases} W\,W^T & if \quad L = 2 \\ W\,W^T\,U^{L-2} & if \quad L = 4,6,8,.. \end{cases}$$

$L \leftarrow L + 2$;

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

Bước 2. Xác định tập láng giềng cho người dùng $i \in U$.

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

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

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

Bước4. Chọn N sản phẩm có mức độ tương tự cao nhất tư vấn cho người dùng i.

Hình 3. Thuật toán UserBased-Graph.

Tại bước 1, thuật toán thực hiện tính toán mức độ tương tự giữa các cặp người dùng dựa vào Định lý 1. Kết quả thực hiện của bước 1 là ma trận $U^L(N \times N)$ phản ánh mức độ tương tự giữa người dùng i và người dùng j trên đồ thị. Tại bước 2, thuật toán tiến hành sắp xếp các giá trị u_{ij}^L ($j \neq i$) theo thứ tự giảm dần của trọng số. Sau đó chọn K người dùng đầu tiên làm tập láng giềng của người dùng i. Tại bước 3, thuật toán dự đoán quan điểm của người dùng i đối với các sản phẩm mới $x \in P \backslash P_i$ bằng cách lấy giá trị trung bình các đánh giá của những người dùng j trong tập láng giềng. Bước 4 chọn K sản phẩm đầu tiên tư vấn cho người dùng i.

Ví dụ với hệ lọc cộng tác được biểu diễn bằng ma trận trọng số C trên Hình 1, ta tính toán được $U^2(3\times3), U^4(3\times3), U^6(3\times3)$ theo công thức (5). Dựa vào đó ta xác định được $L=2$ cho người dùng u_2, $L=4$ cho người dùng u_1 và u_3 và không cần thực hiện với giá trị $L=6$.

$$U^2 = \begin{bmatrix} 1.64 & 0.64 & 0.00 \\ 0.64 & 1.00 & 0.36 \\ 0.00 & 0.36 & 0.52 \end{bmatrix}$$

$$U^4 = \begin{bmatrix} 3.0992 & 1.6896 & 0.2304 \\ 1.6896 & 1.5392 & 0.5472 \\ 0.2304 & 0.5472 & 0.4000 \end{bmatrix}$$

$$U^6 = \begin{bmatrix} 6.164032 & 3.756032 & 0.728064 \\ 3.756032 & 2.817536 & 0.838656 \\ 0.728064 & 0.838656 & 0.404992 \end{bmatrix}$$

III.3. Lọc cộng tác sử dụng độ tương tự giữa cặp sản phẩm trên đồ thị

Do vai trò của người dùng và sản phẩm trong ma trận đánh giá là đối xứng, ta có thể xây dựng phiên bản lọc cộng tác sử dụng độ tương tự giữa các sản phẩm, trong đó độ tương tự được tính toán dựa trên đồ thị theo cách tương tự như trình bầy ở trên.

Gọi $P^L(M \times M)$ là tổng trọng số các đường đi có độ dài L từ đỉnh $x \in P$ đến đỉnh $y \in P$ trên đồ thị G ($x=1, 2,.., M; j=1, 2, .., M$). Vì tổng trọng số mỗi đường đi độ dài L được tính bằng tích của trọng số các cạnh, nên tổng trọng số các đường đi độ dài L từ đỉnh sản phẩm đến đỉnh sản phẩm trên đồ thị G được xác định theo công thức (11).

$$P^L = \begin{cases} W^T.W & if \quad L = 2 \\ W^T.W.P^{L-2} & if \quad L = 4,6,8,.. \end{cases} \qquad (6)$$

Mức độ tương tự giữa các cặp sản phẩm xác định theo (6) cũng phụ thuộc vào độ dài đường đi L từ đỉnh

sản phẩm đến đỉnh sản phẩm trên đồ thị. Do vậy, với mỗi sản phẩm $x \in P$ ta cũng cần xác định giá trị của L để thực hiện tính toán. Định lý 3 dưới đây sẽ cho ta một cách xác định L trong trường hợp đồ thị biểu diễn của lọc cộng tác $G = <V,E>$ liên thông. Đối với các hệ lọc cộng tác có biểu diễn đồ thị $G = <V, E>$ không liên thông chúng tôi sẽ trình bày trong những kết quả nghiên cứu tiếp theo của bài báo.

Định lý 2. Nếu đồ thị biểu diễn cho các hệ lọc cộng tác $G = <V, E>$ liên thông thì luôn luôn tồn tại số tự nhiên chẵn L để $p_{xy}^{L} \neq 0$ với mọi $x, y \in P$. Trong đó, P_{xy}^{L} được xác định theo (6).

Định lý 2 được chứng minh tương tự như Định lý 1. Kết quả này cho phép ta xác định giá trị L *nhỏ nhất* để $p_{xy}^{L} \neq 0$ với mọi $x, y \in P$. Ví dụ với hệ lọc cộng tác được biểu diễn bằng ma trận trọng số C trên Hình 1, ta tính toán được $P^2(4\times4), P^4(4\times4), P^6(4\times4)$ theo (6). Dựa vào đó ta xác định được $L=4$ đối với sản phẩm p_2 và p_3 , $L=6$ đối với sản phẩm p_1 và p_4.

$$P^2 = \begin{bmatrix} 1.00 & 0.00 & 0.80 & 0.00 \\ 0.00 & 0.72 & 0.48 & 0.24 \\ 0.80 & 0.48 & 1.28 & 0.00 \\ 0.00 & 0.24 & 0.00 & 0.16 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} 1.6400 & 0.3840 & 1.8240 & 0.0000 \\ 0.3840 & 0.8064 & 0.9600 & 0.2112 \\ 1.8240 & 0.9600 & 2.5088 & 0.1152 \\ 0.0000 & 0.2112 & 0.1152 & 0.0832 \end{bmatrix}$$

$$P^6 = \begin{bmatrix} 3.099200 & 1.152000 & 3.831040 & 0.092160 \\ 1.152000 & 1.092096 & 1.923072 & 0.227328 \\ 3.831040 & 1.923072 & 5.131265 & 0.248832 \\ 0.092160 & 0.227328 & 0.248832 & 0.064000 \end{bmatrix}$$

Dựa trên kết quả của Định lý 2, chúng tôi đề xuất thuật toán ItemBased-Graph cho lọc cộng tác được mô tả chi tiết trong Hình 4. Tại bước 1, thuật toán thực hiện tính toán mức độ tương tự giữa các cặp sản phẩm dựa vào Định lý 2. Kết quả thực hiện của bước 1 là ma trận $P^L(M\times M)$ phản ánh mức độ tương tự giữa sản phẩm x và sản phẩm y trên đồ thị. Tại bước 2, thuật toán tiến hành sắp xếp các giá trị $p_{xy}^{L}(j \neq i)$ theo thứ tự giảm dần của trọng số. Sau đó chọn K sản phẩm đầu tiên làm tập láng giềng của sản phẩm x. Tại bước 3, thuật toán dự đoán quan điểm của người dùng i đối với các sản phẩm mới $x \in P \backslash P_i$ bằng cách lấy giá trị trung bình các đánh giá của những sản phẩm x trong tập láng giềng. Tại bước 4, thuật toán chọn K sản phẩm có mức độ tương tự cao nhất tư vấn cho người dùng i.

Đầu vào:

- Ma trận trọng số C là biểu diễn đồ thị $G = <V, E>$ cho lọc cộng tác.
- $x \in P$ là sản phẩm cần dự đoán
- K là số lượng sản phẩm của tập láng giềng.

Đầu ra:

- Dự đoán x: $r_{ix} \mid x \in U \setminus U_x$.(quan điểm của người dùng i đối với phẩm mới $x \in P$).

Các bước tiến hành:

Bước 1. Tính toán mức độ tương tự giữa các cặp người dùng:

$L \leftarrow 2$;//Khởi tạo độ dài đường đi ban đầu

Repeat

$$P^L = \begin{cases} W^T.W & if \quad L = 2 \\ W^T.W.P^{L-2} & if \quad L = 4,6,8,.. \end{cases}$$

$L \leftarrow L + 2$;

Until ($p_{xy}^{L} \neq 0$ *với mọi* $y \in (P \setminus x)$);

Bước 2. Xác định tập láng giềng cho sản phẩm $x \in P$.

- Sắp xếp p_{xy}^{L} theo thứ tự giảm dần ($x \neq y$).
- Chọn K sản phẩm $y \in P$ đầu tiên làm tập láng giềng của sản phẩm x (Ký hiệu tập láng giềng của người dùng $x \in P$ là K_x).

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

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

Bước4. Chọn K sản phẩm có mức độ tương tự cao nhất tư vấn cho người dùng i.

Hình 4. Thuật toán ItemBased-Graph.

Trong cả hai trường hợp tính toán độ tương tự giữa người dùng và độ tương tự giữa sản phẩm đều có thể xuất hiện trường hợp giữa hai người dùng hoặc hai sản phẩm không tồn tại đường đi trên đồ thị. Việc xác định các trường hợp để tồn tại đường đi, tức là độ tương tự giữa hai đối tượng khác không, được thực hiện dựa trên định lý sau.

Định lý 3. Điều kiện cần và đủ để $U^L(N \times N)$ xác định theo (5), $P^L(M \times M)$ xác định theo (6), được điền đầy đủ giá trị khác 0 khi và chỉ khi đồ thị biểu diễn cho các hệ lọc cộng tác G = <V, E> liên thông.

Chứng minh (*Điều kiện cần*). Giả sử $U^L(N \times N)$, $P^L(M \times M)$, $W^L(N \times M)$ được điền đầy đủ các giá trị khác 0. Khi đó ta cần chứng tỏ *G* liên thông. Thực vậy, vì $U^L(N \times N)$ được điền đầy đủ giá trị khác 0 nên với mọi *i, j*$\in U$ đều tồn tại ít nhất một đường đi có độ dài *L*. $P^L(M \times M)$ được điền đầy đủ giá trị khác 0 nên với mọi *x, y*$\in P$ đều tồn tại ít nhất một đường đi có độ dài *L*. $W^L(N \times M)$ được điền đầy đủ giá trị khác 0 nên với mọi *i*$\in U$, *x*$\in P$ đều tồn tại ít nhất một đường đi có độ dài *L*. Từ đây ta suy ra giữa hai đỉnh bất kỳ của đồ thị đều tồn tại đường đi. Do vậy đồ thị *G* = <*V, E*> liên thông.

Ngược lại (*điều kiện đủ*): Giả sử *G* = <*V, E*> liên thông, theo Định lý 3 $U^L(N \times N)$ sẽ được điền đầy đủ các giá trị khác 0, theo Định lý 2 $P^L(M \times M)$ sẽ được điền đầy đủ các giá trị khác 0, theo Định lý 1 $W^L(N \times M)$ cũng sẽ được điền đầy đủ các giá trị khác 0.

Trong trường hợp đồ thị không liên thông, các định lý trên cho phép xác định đường đi giữa hai người dùng bất kỳ nếu hai người đó cùng thuộc một thành phần liên thông trên đồ thị đánh giá. Kết luận tương tự đối với các cặp sản phẩm.

IV. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi thực hiện tiến hành thử nghiệm trên bộ dữ liệu đánh giá phim và so sánh với một số phương pháp khác. Phần này sẽ trình bầy chi tiết về thử nghiệm và kết quả.

Dữ liệu: Dữ liệu thử nghiệm là bộ dữ liệu MovieLens [13]. 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 1, 2, 3, 4, 5 được chuyển đổi thành 0.2, 0.4, 0.6, 0.8, 1.0.

Phương pháp thử nghiệm: Sai số dự đoán của các phương pháp được ước lượng bằng độ chính xác (*precision*), độ nhậy (*recall*) và độ đo *F* (*F-Measure*). Độ chính xác, độ nhạy, và độ đo F có giá trị lớn phản ánh mức độ chính xác của thuật toán càng cao [7]. 900 người dùng trong tập MovieLens được lựa chọn ngẫu nhiên làm dữ liệu huấn luyện, 400 người dùng được lựa chọn ngẫu nhiên trong số còn lại để làm tập kiểm tra. Để thử nghiệm khả năng của phương pháp mới đề xuất so với những phương pháp khác trong trường hợp dữ liệu thưa, chúng tôi thay đổi số lượng đánh giá của mỗi người dùng trong tập kiểm tra số lượng đánh giá đã biết lần lượt là *5, 10, 15, 20* sao cho đồ thị biểu diễn của lọc cộng tác vẫn liên thông, các đánh giá còn lại được ẩn đi và được dùng để so sánh với kết quả dự đoán. Phương pháp được sử dụng để đưa ra dự đoán với những đánh giá đã bị ẩn. Kết quả dự đoán của các phương pháp được lấy từ trung bình qua 10 lần thử nghiệm, trong mỗi lần, tập huấn luyện và tập kiểm tra được lựa chọn ngẫu nhiên như trên.

So sánh: Kết quả dự đoán của phương pháp UserBased-Graph, IemBased-Graph được so sánh với phương pháp KNN-UserBased [6,7], Top-N-Item Based [3,4,5] dựa trên độ tương quan Pearson và phương pháp GraphBased [10]. Hai phương pháp đầu là phương pháp k-láng giềng gần nhất trong khi phương pháp thứ ba là phương pháp thuần túy dựa trên đồ thị.

Kết quả: Kết quả thử nghiệm được tóm tắt trong Bảng 3. Các kết quả cho thấy phương pháp Top-N-ItemBased có độ đo F cao hơn so với KNN-UserBased trong một số trường hợp nhưng lại thấp hơn trong một

số trường hợp khác tùy thuộc vào tính chất dữ liệu. Kết quả này nhất quán so với các thử nghiệm đã công bố trước đây.

Phương pháp dựa trên đồ thị do Huang và cộng sự đề xuất cho kết quả tốt hơn hai phương pháp k-láng giềng gần nhất trong cả bốn trường hợp, đặc biệt khi dữ liệu thưa. Cụ thể, với chỉ 5 đánh giá cho một người dùng, GraphBased đạt độ đo F bằng 0.178 trong khi không có phương pháp dựa trên bộ nhớ nào có độ đo F vượt quá 0.139.

Bảng 3. Độ chính xác, độ nhạy và tỷ lệ F ứng với các đánh giá biết trước

Phương pháp	Độ đo	Số đánh giá biết trước trong tập kiểm tra			
		5	10	15	20
Top-N-ItemBased	Độ nhạy	0.108	0.118	0.124	0.251
	Độ chính xác	0.164	0.178	0.211	0.244
	F-Measure	0.130	0.142	0.156	0.247
KNN-UserBased	Độ nhạy	0.112	0.131	0.142	0.149
	Độ chính xác	0.184	0.194	0.214	0.265
	F-Measure	0.139	0.156	0.171	0.191
Graph-Based [10]	Độ nhạy	0.173	0.192	0.213	0.256
	Độ chính xác	0.184	0.246	0.259	0.326
	F-Measure	0.178	0.212	0.234	0.287
ItemBased-Graph	Độ nhạy	0.212	0.238	0.275	0.288
	Độ chính xác	0.287	0.256	0.284	0.473
	F-Measure	0.199	0.245	0.279	0.358
UserBased-Graph	Độ nhạy	0.225	0.244	0.287	0.295
	Độ chính xác	0.288	0.308	0.284	0.477
	F-Measure	0.253	0.272	0.290	0.365

Phương pháp đề xuất, theo đó độ tương tự trên đồ thị được sử dụng để xác định k láng giềng gần nhất, cho kết quả tốt hơn hẳn các phương pháp được so sánh. Cả hai phiên bản sử dụng độ tương tự theo sản phẩm hay theo người dùng đều có độ đo F lớn hơn so với phương pháp sử dụng đồ thị thuần túy. Cụ thể, với 5 đánh giá cho mỗi người dùng, UserBased-Graph và ItemBased-Graph cho độ đo F lần lượt là 0.253 và 0.199, so với 0.178 của phương pháp GraphBased. Kết quả này cũng nhất quán khi tăng số lượng đánh giá (giảm độ thưa thớt dữ liệu). Với 20 đánh giá cho một người dùng, UserBased-Graph, ItemBased-Graph, và GraphBased cho độ đo F lần lượt là 0.365, 0.358, và 0.287. Độ chính xác cao hơn của phương pháp đề xuất so với phương pháp dựa trên đồ thị thuần túy có thể là kết quả của việc sử dụng k láng giềng đã tạo hiệu ứng làm trơn nhờ lấy trung bình đánh giá của người dùng hoặc sản phẩm tương tự. Một yếu tố khác ảnh hưởng tốt tới độ chính xác là việc xác định hợp lý độ dài đường đi sao cho tạo được độ liên thông hợp lý đồng thời không gây nhiễu khi chọn đường đi quá dài.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã trình bày một phương pháp tiếp cận cho lọc cộng tác bằng mô hình đồ thị. Trong đó, phương pháp biểu diễn đồ thị phù hợp với tất cả các bộ dữ liệu hệ thống lọc cộng tác hiện nay. Dựa vào biểu diễn này, các phương pháp lọc cộng tác đều được triển khai dễ dàng trên đồ thị. Phương pháp lọc dựa vào người dùng được xem xét như bài toán tìm kiếm và đánh giá trọng số các đường đi từ đỉnh người dùng đến đỉnh người dùng. Phương pháp lọc dựa vào sản phẩm được xem xét như bài toán tìm kiếm và đánh giá trọng số các đường đi từ đỉnh sản phẩm đến đỉnh sản phẩm. Các đường đi sau đó được sử dụng như độ đo tương tự và được kết hợp với phương pháp k – láng giềng gần nhất để đưa ra dự đoán. Kết quả thử nghiệm cho thấy, phương pháp đề xuất đều cho lại kết quả dự đoán tốt hơn các phương pháp lọc dựa trên độ tương quan trong trường hợp có đầy đủ dữ liệu huấn luyện cũng như trường hợp dữ liệu thưa. Điều đó chứng tỏ, phương pháp tiếp cận cho lọc cộng tác bằng mô hình đồ thị cho phép ta khai thác được các mối quan hệ gián tiếp giữa tập người dùng và tập sản phẩm vào quá trình dự đoán. Việc kết hợp quan hệ gián tiếp với phương pháp dựa trên bộ nhớ truyền thống cho kết quả tốt hơn khi sử dụng từng phương pháp riêng rẽ.

TÀI LIỆU THAM KHẢO

[1] Y. KOREN, R. BELL, *Advances in collaborative filtering. Recommender systems handbook*. Springer, 2011.

[2] G. ADOMAVICIUS, A. TUZHILIN, "*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] B. SARWAR, G. KARYPIS, J. KONSTAN, AND J. RIEDL, "*Item-Based Collaborative Filtering Recommendation Algorithms*", pp. 285-295, 2001.

[4] M. DESHPANDE, G. KARYPIS,"Item-Based *Top-N Recommendation Algorithms"*, ACM Transactions on Information Systems. Volume 22, Issue 1, pp. 143 - 177, 2004.

[5] J. WEN, W. ZHOU, *"An Improved Item-based Collaborative Filtering Algorithm Based on Clustering Method", Journal of Computational Information Systems 8*: pp. 571-578, 2012.

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

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

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

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

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

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

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

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

[14] G. XUE, C. LIN, Q. YANG, W. XI, H. ZENG, Y. YU, AND Z. CHEN. *Scalable collaborative filtering using cluster-based smoothing*. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). New York, USA, 114-121.

[15] 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.

[16] L. BALTRUNAS, F. RICCI, *Experimental evaluation of context-dependent collaborative filtering using item splitting*. User modeling and user-adapted interactions (2013). Springer.

Nhận bài ngày: 12/06/2013

SƠ LƯỢC VỀ TÁC GIẢ

NGUYỄN DUY PHƯƠNG

Sinh ngày 20/02/1965 tại Hà Nội.

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

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

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

Email: phuong.ptit@yahoo.com

Điện thoại : 0913575442

TỪ MINH PHƯƠNG

Sinh ngày 13/01/1971 tại Hà Nội.

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

Hiện là Phó Giáo sư, Trưởng khoa CNTT - Học viện CN Bưu chính Viễn thông.

Hướng nghiên cứu: trí tuệ nhân tạo, học máy, tin sinh học