Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00060

MỘT TIẾP CẬN ĐA QUAN HỆ CHO HỆ THỐNG GỢI Ý

Nguyễn Thái Nghe, Mai Nhựt Tự, Nguyễn Hữu Hòa

Khoa Công nghệ thông tin & Truyền Thông, Trường Đại học Cần Thơ

{ntnghe,nhhoa}@ctu.edu.vn, mntu.it@gmail.com

TÓM TẮT *— Kỹ thuật phân rã ma trận (matrix factorization - MF) là một kỹ thuật được sử dụng phổ biến trong hệ thống gợi ý (Recommender Systems – RS). Hiện nay đã có rất nhiều thuật toán biến thể và hướng tiếp cận được phát triển dựa trên kỹ thuật này, như Biased matrix factorization, Non-negative matrix factorization, phân rã ma trận đa quan hệ (multi-relational matrix factorization - MRMF),... Bài viết này đề xuất một hướng tiếp cận đa quan hệ khác cho hệ thống gợi ý, từ đó xây dựng thuật toán cho hướng tiếp cận đã đề xuất. Thực nghiệm trên các tập dữ liệu chuẩn trong lĩnh vực gợi ý dùng độ đo RMSE (Root Mean Squared Error) cho thấy hướng tiếp cận đã đề xuất cho kết quả rất khả quan.*

Từ khoá *— Phân rã ma trận; phân rã ma trận đa quan hệ; hệ thống gợi ý.*

I. GIỚI THIỆU

Hệ thống gợi ý (Recommender Systems - RS) đã được ứng dụng rất rộng rãi trong các hệ thống thông tin thuộc nhiều lĩnh vực khác nhau, nó giúp giải quyết được vấn đề quá tải về thông tin và giúp lựa chọn thông tin một cách nhanh chóng bằng cách trình bày nội dung gợi ý phù hợp với từng người dùng. Để cung cấp cho người dùng những thông tin gợi ý hiệu quả thì mỗi hệ thống gợi ý cần có một mô hình gợi ý có thể khai thác tốt được dữ liệu đã thu thập để đưa ra các gợi ý phù hợp cho từng người dùng, do đó việc lựa chọn thuật toán xây dựng mô hình gợi ý là rất quan trọng.

Trong RS đã có rất nhiều giải thuật được đề xuất, tuy nhiên ta có thể gom chúng vào trong ba nhóm chính (xem thêm trong [1], [2])

- Nhóm giải thuật lọc trên nội dung (Content-based Filtering): Thực hiện việc gợi ý các mục dữ liệu (item) dựa vào hồ sơ (profiles) của người dùng hoặc dựa vào thuộc tính (attributes) của những item tương tự như item mà người dùng đã chọn trong quá khứ.
- Nhóm giải thuật lọc cộng tác (Collaborative Filtering): Các giải thuật trong nhóm này chủ yếu dựa trên các kỹ thuật: phương pháp láng giềng (Neighborhood-based) là dựa vào dữ liệu quá khứ của những người dùng "tương tự" (user-based approach) hoặc dựa trên dữ liệu quá khứ của những item "tương tự" (item-based apprach); dựa trên mô hình (Model-based): nhóm này xây dựng các mô hình dự đoán dựa trên dữ liệu thu thập được trong quá khứ.
- Nhóm kết hợp cả 2 cách trên.

Trong nhóm giải thuật lọc cộng tác dựa trên mô hình thì kỹ thuật phân rã ma trận (matrix factorization - MF) là một trong những phương pháp thành công nhất hiện nay (state-of-the-art) trong lĩnh vực dự đoán xếp hạng của RS [3], [4]. Tuy nhiên, đa số các giải thuật thuộc nhóm MF chỉ tập trung khai thác thông tin của một mối quan hệ đơn giữa người dùng (user) và mục dữ liệu (item) chẳng hạn quan hệ đánh giá (rating), do đó các giải thuật chưa tận dụng được hết các thông tin liên quan từ các mối quan hệ khác của user và item. Để tận dụng hết các thông tin, người ta đã đề xuất phương pháp phân rã ma trận đa quan hệ (multi-relational matrix factorization - MRMF) như trong [5], [6], mặc dù vậy trong các nghiên cứu này, công thức dùng cho dự đoán vẫn chưa bao gồm hết thông tin từ các ma trận nhân tố tiềm ẩn (sẽ được phân tích sau).

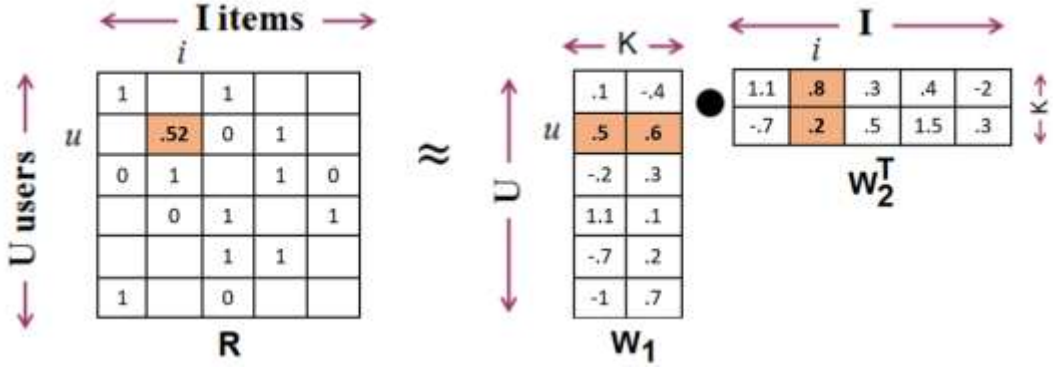
Trong bài viết này, chúng tôi sẽ đề xuất một hướng tiếp cận đa quan hệ cho hệ thống gợi ý cho phép tận dụng được thông tin từ nhiều mối quan hệ khác nhau của user và item trong quá trình xây dựng mô hình và đưa ra gợi ý, từ đó xây dựng thuật toán cho hướng tiếp cận đã đề xuất. Chúng tôi đã thực nghiệm trên các tập dữ liệu chuẩn trong lĩnh vực Hệ thống gợi ý và cả lĩnh vực Hệ trợ giảng thông minh để đánh giá độ chính xác của mô hình thông qua chỉ số RMSE (Root Mean Squared Error). Kết quả cho thấy hướng tiếp cận được đề xuất rất có thể giúp cải thiện độ chính xác.

II. KỸ THUẬT PHÂN RÃ MA TRẬN ĐA QUAN HỆ VÀ NHỮNG NGHIÊN CỨU LIÊN QUAN

Trước tiên chúng tôi tóm tắt ngắn gọn kỹ thuật phân rã ma trận trên quan hệ đơn (MF) (xem thêm trong bài viết [4]) và kỹ thuật phân rã ma trận đa quan hệ (MRMF) (xem thêm trong bài viết [5], [11]) để làm cơ sở cho việc đề xuất một hướng tiếp cận đa quan hệ mới.

A. Kỹ thuật phân rã ma trận (Matrix Factorization - MF)

Kỹ thuật phân rã ma trận là việc chia một ma trận lớn $\mathbf{R}$ thành hai ma trận $\mathbf{W_1}$ và $\mathbf{W_2}$ có kích thước nhỏ hơn rất nhiều so với ma trận $\mathbf{R}$, sao cho $\mathbf{R}$ có thể được xây dựng lại từ hai ma trận nhỏ hơn này càng chính xác càng tốt [4], nghĩa là $\mathbf{R} \approx \mathbf{W_1}\mathbf{W_2^T}$ như minh hoạ trong Hình 1.



Hình 1. Minh họa kỹ thuật phân rã ma trận

$\mathbf{W_1} \in \mathbb{R}^{|U| \times K}$ là một ma trận mà ở đó mỗi dòng *u* là một véctơ bao gồm *K* nhân tố tiềm ẩn (latent factors) mô tả cho người dùng *u*, và $\mathbf{W_2} \in \mathbb{R}^{|I| \times K}$ là một ma trận mà ở đó mỗi dòng *i* là một véctơ bao gồm *K* nhân tố tiềm ẩn mô tả cho mục dữ liệu *i*.

Gọi $\boldsymbol{w}_{1_{uk}}$ và $\boldsymbol{w}_{2_{ik}}$ là các phần tử tương ứng của hai ma trận $\mathbf{W_1}$ và $\mathbf{W_2}$ hay $\mathbf{w}_{1_u}$và $\mathbf{w}_{2_i}$ là các véctơ bao gồm K nhân tố tiềm ẩn mô tả cho người dùng *u* và item *i* khi đó xếp hạng của user *u* trên item *i* được dự đoán bởi công thức:

$$\hat{r}_{ui} = \sum_{k=1}^{K} \mathbf{w}_{1_{uk}} \mathbf{w}_{2_{ik}} = \mathbf{w}_{1_u} \mathbf{w}_{2_i}^T \tag{1}$$

$\mathbf{W_1}$ và $\mathbf{W_2}$ là các tham số mô hình (còn gọi là các ma trận nhân tố tiềm ẩn) mà chúng ta cần phải xác định bằng cách tối ưu hóa hàm mục tiêu (2) theo một điều kiện nào đó, chẳng hạn RMSE (root mean squared error).

$$\text{RMSE} = \sqrt{\frac{1}{|D^{\text{test}}|} \sum_{u,i,r \in D^{\text{test}}} \left(r_{ui} - \hat{r}_{ui}\right)^2}$$

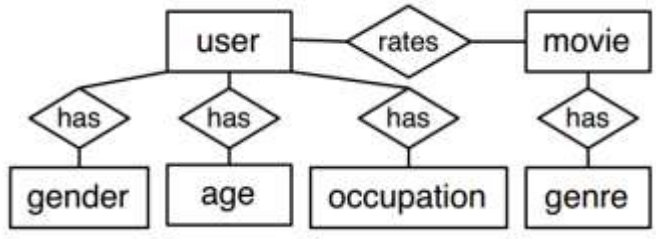
Hàm mục tiêu dùng cho việc tối ưu hoá để xác định $\mathbf{W_1}$ và $\mathbf{W_2}$ được trình bày như sau:

$$O^{MF} = \sum_{(u,i) \in \mathbf{R}} \left(\mathbf{R}_{ui} - \mathbf{w}_{1_u} \mathbf{w}_{2_i}^T\right)^2 + \lambda(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W_2}\|_F^2) \tag{2}$$

Với λ là hệ số chính tắc hóa ($0 \le \lambda < 1$) và $\|\cdot\|_F^2$ là chuẩn Frobenius[1]. Đại lượng $\lambda \cdot (\|\mathbf{W}_1\|_F^2 + \|\mathbf{W_2}\|_F^2)$ được dùng để ngăn ngừa sự quá khớp (over-fitting) [9]. Để tối ưu hóa hàm mục tiêu, người ta có thể dùng phương pháp giảm dốc ngẫu nhiên (stochastic gradient descent - SGD) [8].

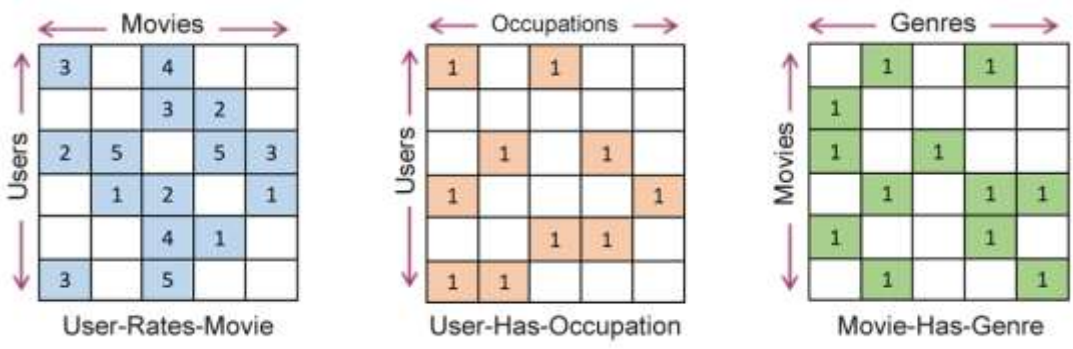
B. Kỹ thuật phân rã ma trận đa quan hệ (Multi-Relational Matrix Factorization - MRMF)

Trong phần trước, chúng tôi đã trình bày tóm tắt kỹ thuật phân rã ma trận, kỹ thuật này chỉ sử dụng được một quan hệ đơn giữa hai thực thể (ví dụ như quan hệ "rates" giữa hai thực thể "user" và "movie" trong Hình 2). Kỹ thuật phân rã ma trận đa quan hệ (MRMF) [5] là trường hợp tổng quát hơn của MF, với MRMF chúng ta có thể sử dụng nhiều hơn một quan hệ và nhiều hơn hai thực thể trong mô hình.



Hình 2. Ví dụ mô hình thực thể kết hợp (ERD) của tập dữ liệu MovieLens

[1] https://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm



Hình 3. Ví dụ biểu diễn thông tin từ mô hình ERD trong Hình 2 vào các ma trận

Nếu gọi $\{\mathbf{E}_1, \mathbf{E}_2,\ldots, \mathbf{E}_N\}$ là một tập N thực thể, $\{\mathbf{R}_1, \mathbf{R}_2,\ldots, \mathbf{R}_M\}$ là một tập M quan hệ nhị phân và $\mathbf{R}_r = \{(E_{1_r}; E_{2_r})\}$ với ($r = 1 \ldots M$), khi đó hàm mục tiêu cần phải tối ưu của MRMF sẽ là:

$$O^{MRMF} = \sum_{r=1}^{M} \sum_{(u,i) \in \mathbf{R}_r} \left((\mathbf{R}_r)_{ui} - \mathbf{w}_{r_1 u} \mathbf{w}_{r_2 i}^T\right)^2 + \lambda \left(\sum_{j=1}^{N} \|\mathbf{W}_j\|_F^2 \right) \quad (3)$$

Trong đó N là số thực thể, M là số quan hệ, $\{\mathbf{W}_j\}_{j = 1 \ldots N}$ là các ma trận nhân tố tiềm ẩn của N thực thể, $(\mathbf{R}_r)_{ui}$ là giá trị thực của user u trên item i trong quan hệ thứ r, $\mathbf{w}_{r_1 u} \mathbf{w}_{r_2 i}^T$ là giá trị dự đoán của user u trên item i trong quan hệ thứ r, $\lambda \left(\sum_{j=1}^{N} \|\mathbf{W}_j\|_F^2 \right)$ là đại lượng chính tắc hóa tương tự như MF. Hàm mục tiêu (3) vẫn được tối ưu bằng phương pháp SGD, tức là các tham số **w** tương ứng sẽ được cập nhật theo công thức:

$$\mathbf{w}_{r_1 u}^{new} = \boldsymbol{w}_{r_1 u}^{old} - \beta \left(\frac{\partial O^{MRMF}}{\partial \boldsymbol{w}_{r_1 u}^{old}} \right) \quad (4)$$

$$\mathbf{w}_{r_2 i}^{new} = \boldsymbol{w}_{r_2 i}^{old} - \beta \left(\frac{\partial O^{MRMF}}{\partial \boldsymbol{w}_{r_2 i}^{old}} \right) \quad (5)$$

Với β là tốc độ học (learning rate, $0 < \beta < 1$). Giá trị của $\frac{\partial O^{MRMF}}{\partial \mathbf{w}_{r_1 u}}$ và $\frac{\partial O^{MRMF}}{\partial \mathbf{w}_{r_2 i}}$ được xác định bởi công thức:

$$\frac{\partial O^{MRMF}}{\partial \mathbf{w}_{r_1 u}} = -2 \left((\mathbf{R}_r)_{ui} - \mathbf{w}_{r_1 u} \mathbf{w}_{r_2 i}^T\right) \mathbf{w}_{r_2 i} + \lambda \, \mathbf{w}_{r_1 u} \quad (6)$$

$$\frac{\partial O^{MRMF}}{\partial \mathbf{w}_{r_2 i}} = -2 \left((\mathbf{R}_r)_{ui} - \mathbf{w}_{r_1 u} \mathbf{w}_{r_2 i}^T\right) \mathbf{w}_{r_1 u} + \lambda \, \mathbf{w}_{r_2 i} \quad (7)$$

Bên cạnh đó, nghiên cứu [6] cũng đã ứng dụng kỹ thuật MRMF vào vấn đề dự đoán năng lực của sinh viên. Nhóm tác giả đã tận dụng khả năng mở rộng có thể xử lý nhiều quan hệ và nhiều thực thể của kỹ thuật MRMF để tận dụng được tối đa các thông tin của sinh viên và các thông tin của công việc mà sinh viên cần giải quyết, từ đó giúp cho mô hình dự đoán có độ chính xác cao. Ngoài ra, tác giả còn giới thiệu thêm một biến thể của MRMF được gọi là kỹ thuật phân rã ma trận đa quan hệ với trọng số (Weighted Multi-Relational Matrix Factorization - WMRMF), kỹ thuật này tương tự như MRMF nhưng nó cho phép gán trọng số (Θ_r) cho từng quan hệ nhằm phân biệt mức độ quan trọng của các quan hệ trong quá trình xây dựng mô hình dự đoán như trình bày trong công thức (8)

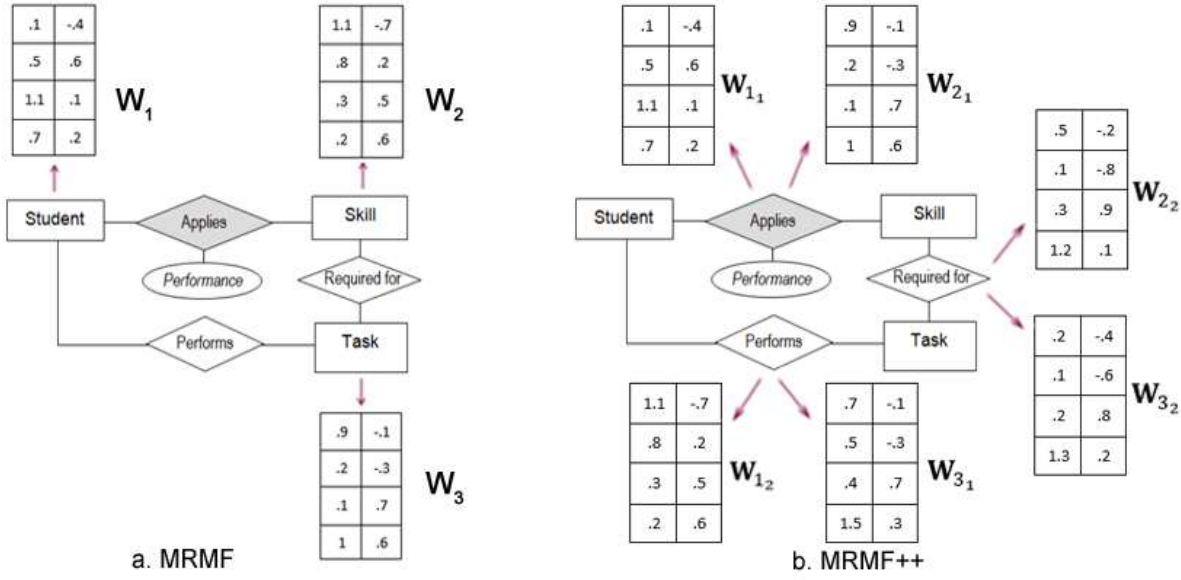
$$O^{WMRMF} = \sum_{r=1}^{M} \Theta_r \sum_{(u,i) \in \mathbf{R}_r} \left((\mathbf{R}_r)_{ui} - \mathbf{w}_{r_1 u} \mathbf{w}_{r_2 i}^T\right)^2 + \lambda \left(\sum_{j=1}^{N} \|\mathbf{W}_j\|_F^2 \right) \quad (8)$$

Tuy quá trình tối ưu hóa hàm mục tiêu để tìm ra các tham số mô hình của kỹ thuật MRMF/WMRMF đã có thể tận dụng được nhiều thông tin hơn MF nhưng kết quả dự đoán của MRMF/WMRMF vẫn được tính bằng công thức (1) như ta thấy trong hàm mục tiêu (3) (8). Do đó, MRMF/WMRMF vẫn chưa khai thác được hết thông tin từ các ma trận nhân tố tiềm ẩn trong việc dự đoán kết quả. Trong bài viết này, chúng tôi đề xuất một hướng MRMF khác có thể khắc phục được nhược điểm trong những nghiên cứu trước đây như đã trình bày trên.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Với mục tiêu tận dụng hết thông tin từ các ma trận nhân tố tiềm ẩn (latent factor matrices) tìm được, chúng tôi đề xuất một hướng tiếp cận đa quan hệ khác có thể tích hợp được thông tin từ các ma trận có liên quan đến user và item vào công thức dự đoán (1) thay vì chỉ sử dụng thông tin của 2 ma trận nhân tố tiềm ẩn đại diện cho user và item như

MRMF. Do đó, số lượng tham số mô hình tìm được của hướng tiếp cận mới sẽ khác với MRMF, chúng tôi gọi hướng tiếp cận mới này là **MRMF++** (Multi-Relational Matrix Factorization Plus Plus).



Hình 4. So sánh tham số mô hình của MRMF và MRMF++

Hình 4 cho thấy sự khác nhau về số lượng tham số mô hình của MRMF và MRMF++. Nếu gọi $\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_N\}$ là một tập N thực thể, $\{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M\}$ là một tập M quan hệ nhị phân và $\mathbf{R}_r = \{(E_{1_r}; E_{2_r})\}$ với ($r = 1 \ldots M$) thì với MRMF số lượng tham số mô hình tìm được sẽ là N ma trận nhân tố tiềm ẩn, với MRMF++ số lượng tham số mô hình tìm được sẽ là $2M$ ma trận nhân tố tiềm ẩn. Dựa trên ý tưởng tận dụng hết thông tin từ các ma trận nhân tố tiềm ẩn trong việc dự đoán, khi đó công thức dự đoán của MRMF++ sẽ là:

$$\hat{r}_{ui} = \left(\sum_{x=1}^{P} w_{1_x}\right)_u \left(\sum_{y=1}^{Q} w_{2_y}\right)_i^T = \sum_{k=1}^{K}\left(\left(\sum_{x=1}^{P} w_{1_x}\right)_{uk}\left(\sum_{y=1}^{Q} w_{2_y}\right)_{ik}\right) \tag{9}$$

Với P và Q tương ứng là số ma trận nhân tố tiềm ẩn của u và i trong toàn bộ mô hình thực thể kết hợp; $\boldsymbol{w}_{1_x}$ là ma trận thứ x trong số P ma trận nhân tố tiềm ẩn của u; $\boldsymbol{w}_{2_y}$ là ma trận thứ y trong số Q ma trận nhân tố tiềm ẩn của i. Để công thức được ngắn gọn ta đặt:

$$X = \sum_{x=1}^{P} w_{1_x}\,;\; Y = \sum_{y=1}^{Q} w_{2_y}$$

Khi đó công thức (9) được viết lại như sau:

$$\hat{r}_{ui} = X_u Y_i^T \tag{10}$$

Tương tự như kỹ thuật MRMF, các tham số mô hình của MRMF++ cũng được tìm bằng cách tối ưu hóa hàm mục tiêu theo một điều kiện nào đó, trong bài viết này chúng tôi sử dụng độ lỗi RMSE. Hàm mục tiêu cần tối ưu là:

$$O^{MRMF++} = \sum_{r=1}^{M}\sum_{(u,i)\in \boldsymbol{R}_r}\left((\boldsymbol{R}_r)_{ui} - X_{ru}{Y_{ri}}^T\right)^2 + \lambda\left(\sum_{j=1}^{2M}\|W_j\|_F^2\right) \tag{11}$$

Trong đó M là số quan hệ, $\{\mathbf{W}_j\}_{j = 1 \ldots 2M}$ là các ma trận nhân tố tiềm ẩn của M quan hệ, $(\mathbf{R}_r)_{ui}$ là giá trị thực của user u trên item i trong quan hệ thứ r, $X_{ru}Y_{ri}^T$ là giá trị dự đoán của user u trên item i trong quan hệ thứ r, $\lambda\left(\sum_{j=1}^{2M}\|\mathbf{W}_j\|_F^2\right)$ là đại lượng chính tắc hóa (regularization). Quá trình tối ưu hóa hàm mục tiêu của MRMF++ được thực hiện bằng phương pháp giảm dốc ngẫu nhiên (stochastic gradient descent - SGD), chẳng hạn ở lần lặp thứ n ta cập nhật các tham số thông qua công thức (12) và (13):

$$X_{ru}^n = X_{ru}^{n-1} - \beta\left(\frac{\partial O^{MRMF++}}{\partial X_{ru}^{n-1}}\right) \tag{12}$$

$$Y_{ri}^n = Y_{ri}^{n-1} - \beta\left(\frac{\partial O^{MRMF++}}{\partial Y_{ri}^{n-1}}\right) \tag{13}$$

Với β là tốc độ học (learning rate, 0 < β < *1*). Giá trị của $\frac{\partial O^{MRMF++}}{\partial X_{ru}}$ và $\frac{\partial O^{MRMF++}}{\partial Y_{ri}}$ được xác định bởi công thức:

$$\left(\frac{\partial O^{MRMF++}}{\partial X_{ru}}\right) = \lambda X_{ru} - 2\left((\boldsymbol{R}_r)_{ui} - X_{ru}{Y_{ri}}^T\right)Y_{ri}^n \quad (14)$$

$$\left(\frac{\partial O^{MRMF++}}{\partial Y_{ri}^n}\right) = \lambda Y_{ri}^n - 2\left((\boldsymbol{R}_r)_{ui} - X_{ru}{Y_{ri}}^T\right)X_{ru} \quad (15)$$

Quá trình tối ưu được tóm tắt trong thủ tục **LearnMRMF++**. Trước tiên, chúng ta cần khởi tạo giá trị của các tham số một cách ngẫu nhiên theo chuẩn phân phối $\mathcal{N}(\mu, \sigma^2)$, ví dụ giá trị trung bình $\mu = 0$, độ lệch chuẩn $\sigma^2 = 0.01$. Khi điều kiện chưa thỏa mãn, chẳng hạn như đạt đến số lần lặp tối đa hoặc tới điểm hội tụ thì các tham số vẫn còn cập nhật (converging: $O_{Iter_{(n-1)}}^{MRMF++} - O_{Iter_n}^{MRMF++} < \epsilon$).

1: **procedure** LEARNMRMF++(**E**$_1$, . . . , **E**$_N$: Thực thể; **R**$_1$, . . . , **R**$_M$: Quan hệ; λ: hằng số chính tắc hóa; β: Tốc độ học; *K*: Số nhân tố tiềm ẩn; Điều kiện dừng)
2: **for** $j \leftarrow 1 \ldots 2M$ **do**
3: $\mathbf{W}_j \leftarrow$ Rút ngẫu nhiên từ $\mathcal{N}(\mu, \sigma^2)$
4: **end for**
5: **while** (Điều kiện dừng chưa thỏa) **do**
6: **for** mỗi quan hệ $\boldsymbol{R}_r = \{(E_{1_r}; E_{2_r})\}$ in {**R**$_1$, . . . , **R**$_M$} **do**
10: **for** $l \leftarrow 1 \ldots |\mathbf{R}_r|$ **do**
11: Lấy ngẫu nhiên bộ (u, i) trong **R**$_r$
12: $X_{ru} \leftarrow X_{ru} - \beta\left(\frac{\partial O^{MRMF++}}{\partial X_{ru}}\right)$
13: $Y_{ri} \leftarrow Y_{ri} - \beta\left(\frac{\partial O^{MRMF++}}{\partial Y_{ri}}\right)$
14: **end for**
15: **end for**
16: **end while**
17: **return** $\{\mathbf{W}_j\}_{j=1...2M}$
18: **end procedure**

Sau quá trình tối ưu, ta nhận được các tham số $\{\mathbf{W}_j\}_{j=1...2M}$, khi đó, chúng ta có thể dự đoán kết quả xếp hạng cho người dùng u trên item i thông qua công thức:

$$\hat{r}_{ui} = \left(\sum_{x=1}^{P} w_{1_x}\right)_u \left(\sum_{y=1}^{Q} w_{2_y}\right)_i^T = \sum_{k=1}^{K}\left(\left(\sum_{x=1}^{P} w_{1_x}\right)_{uk}\left(\sum_{y=1}^{Q} w_{2_y}\right)_{ik}\right)$$

IV. KẾT QUẢ THỰC NGHIỆM

A. Dữ liệu

Để thực nghiệm chúng tôi sử dụng các tập dữ liệu từ hai lĩnh vực khác nhau là trong giải trí và trong giáo dục. Cụ thể, tập dữ liệu Movielens 100k được thu thập bởi nhóm nghiên cứu GroupLens[2]. Dữ liệu này được trích ra từ hệ thống gợi ý phim cho người dùng, tập dữ liệu này có 100.000 đánh giá được thực hiện bởi 943 người dùng trên số lượng 1.682 phim, mỗi người dùng có đánh giá ít nhất 20 phim và đánh giá được gán 1 (tệ) đến 5 (tuyệt vời).

Tập dữ liệu Assistments-2009-2010 (Assistments) trích từ hệ thống ASSISTments[3] [9]. Dữ liệu này có nguồn gốc từ hệ thống trợ giảng thông minh, trong đó sinh viên sẽ giải quyết các bài tập, câu hỏi,... (gọi chung là công việc) và đạt được kết quả tương ứng. Từ kết quả đã có, hệ thống sẽ dự đoán công việc kế tiếp sinh viên có khả năng thực hiện đúng hay sai để đưa ra những phản hồi, gợi ý thích hợp. Dữ liệu này có thể được ánh xạ tương ứng qua các khái niệm trong RS như: sinh viên → user; công việc → item; và kết quả → rating. Thông tin của 2 tập dữ liệu được mô tả trong Bảng 1.

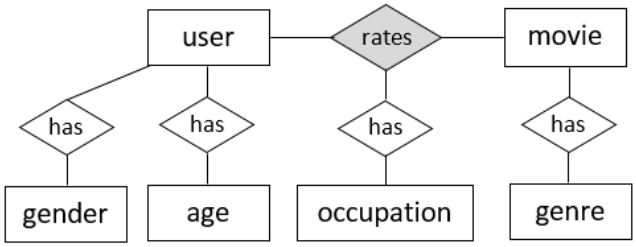
Bảng 1. Thông tin về dữ liệu sử dụng trong thực nghiệm

Tập dữ liệu	Số lượng user	Số lượng item	Số lượng rating
Movielens 100k	943	1,682	100,000
Assistments	8,519	35,798	1,011,079

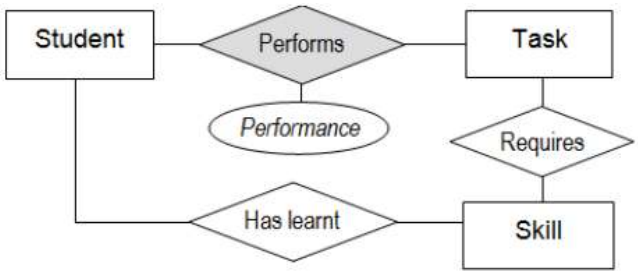
[2] http://www.grouplens.org/node/73.
[3] http://teacherwiki.assistment.org/wiki/Assistments_2009-2010_Full_Dataset.

B. Mô hình thực thể kết hợp (ERD)

Để sử dụng được MRMF và MRMF++ thì cần phải cung cấp danh sách các thực thể (entities) và các quan hệ (relations) để làm tham số đầu vào nên các tập dữ liệu sử dụng trong thực nghiệm cần được tiền xử lý. Ở đây, các ERD được minh hoạ như Hình 5 cho tập Movielens và Hình 6 cho tập Assistments.



Hình 5. Mô hình ERD cho tập Movielens 100k



Hình 6. Mô hình ERD cho tập Assistments

Trong mỗi ERD, quan hệ chính cần dự đoán được tô màu xám, các quan hệ còn lại dùng để bổ sung thông tin trong quá trình xây dựng mô hình dự đoán.

C. Xác định siêu tham số (hyper-parameters)

Các siêu tham số (hyper-parameters) của MF, MRMF, MRMF++ như số lần lặp (Iter), số nhân tố tiềm ẩn K, tốc độ học β, và hệ số hình tắc hóa λ được xác định bằng phương pháp tìm kiếm siêu tham số (hyper-parameter search) [10]. Tuy nhiên, do các tập dữ liệu khá lớn nên việc tìm kiếm bằng vét cạn sẽ mất nhiều thời gian, nên chúng tôi chỉ thực hiện việc tìm kiếm thô cho các phương pháp này: ví dụ: Iter ∈ (50, 100, . . . , 1000), K ∈ (2^4, 2^5, . . . ,2^8), β ∈ (10^{-4}, 10^{-3}, 10^{-2}, $5 \cdot 10^{-5}$, $5 \cdot 10^{-4}$, $5 \cdot 10^{-3}$), λ ∈ ($15 \cdot 10^{-4}$,$15 \cdot 10^{-3}$,$55 \cdot 10^{-5}$,$55 \cdot 10^{-4}$, $55 \cdot 10^{-3}$).

D. Kết quả thực nghiệm

Hình 7 trình bày kết quả thực nghiệm đánh giá bằng RMSE trên tập dữ liệu thu thập từ hệ thống gợi ý phim (Movielens 100k) và tập dữ liệu trích ra từ hệ thống trợ giảng thông minh (Assistments). Kết quả cho thấy rằng kỹ thuật MFMF++ tận dụng được thông tin từ nhiều mối quan hệ khác nhau, cho kết quả cải thiện hơn so với những phương pháp đã có. Những kết quả này cũng đồng nhất với kết quả của các nghiên cứu trước đây trong lĩnh vực hệ thống gợi ý [6], [13], điều này chứng tỏ rằng kỹ thuật MFMR++ là một hướng tiếp cận khả thi cho những dữ liệu đa quan hệ.

Hình 7. Kết quả so sánh RMSE trên tập Movielens 100k và Assistments

V. KẾT LUẬN

Trong bài viết này chúng tôi đã giới thiệu một hướng tiếp cận đa quan hệ mới cho hệ thống gợi ý (Multi-Relational Matrix Factorization Plus Plus – MRMF++). Với MRMF++, mô hình có thể tận dụng được tất cả các mối quan hệ có thể có giữa user và item để dự đoán, đồng thời cũng tận dụng được hết thông tin từ các tham số mô hình để tính toán giá trị dự đoán, do đó có khả năng cải thiện kết quả. Kết quả thực nghiệm cho thấy MRMF++ hoạt động tốt trên cả dữ liệu giải trí và giáo dục, chứng tỏ đây là một hướng tiếp cận khả thi. Tuy nhiên, MRMF++ vẫn còn hạn chế là thời gian huấn luyện mô hình khá chậm so với MRMF, nguyên nhân là do số lượng tham số mô hình cần tìm MRMF++ nhiều hơn dẫn đến quá trình tối ưu hóa hàm mục tiêu cũng mất nhiều thời gian hơn. Trong tương lai, chúng tôi sẽ tiếp tục thực nghiệm trên nhiều tập dữ liệu khác để cũng cố thêm kết quả của phương pháp đề xuất và nghiên cứu giải pháp để cải thiện tốc độ huấn luyện mô hình cho MRMF++.

TÀI LIỆU THAM KHẢO

[1] Ricci, F., Rokach, L., Shapira, B. & Kantor, P.B., eds. (2011). Recommender Systems Handbook. Springer.

[2] Su, X. & Khoshgoftaar, T.M. (2009). A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, 4:1-4:19.

[3] Bell, R. M., & Koren, Y. (2007). Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *Proceedings of the 7th IEEE International Conference on Data Mining* (ICDM 2007), (pp. 43-52). Washington, USA. IEEE CS.

[4] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer Society Press*, 42(8), 30-37.

[5] Lippert, C., Weber, S. H., Huang, Y., Tresp, V., Schubert, M., & Kriegel, H. P. (2008, December). Relation prediction in multi-relational domains using matrix factorization. In *Proceedings of the NIPS 2008 Workshop: Structured Input-Structured Output, Vancouver, Canada*.

[6] Thai-Nghe, N., & Schmidt-Thieme, L. (2015, October). Multi-relational Factorization Models for Student Modeling in Intelligent Tutoring Systems. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on* (pp. 61-66). IEEE.

[7] Bottou, L. (2004). Stochastic learning. In *Advanced lectures on machine learning* (pp. 146-168). Springer Berlin Heidelberg.

[8] Koren, Y. (2010). Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(1), 1.

[9] Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243-266.

[10] Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010, July). Cost-sensitive learning methods for imbalanced data. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1-8). IEEE.

[11] Singh, A. P., & Gordon, G. J. (2008). Relational learning via collective matrix factorization. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, ser. KDD '08. New York, NY, USA: ACM, pp. 650–658.

[12] Cen, H., Koedinger, K., & Junker, B. (2006, June). Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent tutoring systems* (pp. 164-175). Springer Berlin Heidelberg.

[13] Nguyễn Thái Nghe. (2012). Kỹ thuật phân rã ma trận trong xây dựng hệ thống gợi ý. *Kỷ yếu Hội thảo Công nghệ thông tin 2012*. Trường Đại học Đà Lạt, trang 68-77.

[14] Nguyễn Hùng Dũng, Nguyễn Thái Nghe. (2014). Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác. Tạp chí Khoa học - Trường Đại học Cần Thơ, số 31a (2014), trang 36-51. ISSN: 1859-2333.

[15] London, B., Rekatsinas, T., Huang, B., & Getoor, L. (2013). Multi-relational learning using weighted tensor decomposition with modular loss. *arXiv preprint arXiv:1303.1733*.

[16] Jenatton, R., Roux, N. L., Bordes, A., & Obozinski, G. R. (2012). A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems* (pp. 3167-3175).

[17] Nickel, M., Jiang, X., & Tresp, V. (2014). Reducing the rank in relational factorization models by including observable patterns. In *Advances in Neural Information Processing Systems* (pp. 1179-1187).

A MULTI-RELATIONAL APPROACH FOR RECOMMENDER SYSTEMS

Nguyen Thai Nghe, Mai Nhut Tu, Nguyen Huu Hoa

ABSTRACT *— Matrix factorization technique is used most commonly in recommender system. Currently, there are many variations algorithms and approaches developed based on this technique e.g., Biased matrix factorization, Non-negative matrix factorization, multi-relational matrix factorization, etc. The paper proposes a multi-relational approach for recommender systems, thereby building algorithm for the proposed approach, experiments on the standard dataset of recommendation fields (e.g., MovieLens, Assistment, Algebra, etc) and evaluates accuracy through RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error).*

Keywords *— Multi-relational matrix factorization, matrix factorization, recommender systems.*