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.2016.00045

MỘT GIẢI PHÁP XỬ LÝ VẤN ĐỀ NGƯỜI DÙNG MỚI TRONG HỆ THỐNG GỢI Ý

Đinh Thế An Huy[1], Châu Lê Sa Lin[1], Nguyễn Hữu Hòa[2], Nguyễn Thái Nghe[2]

[1]Khoa Công nghệ Thông tin & Truyền Thông, Trường Cao đẳng Kinh tế - Kỹ thuật Cần Thơ
[2]Khoa Công nghệ Thông tin & Truyền Thông, Trường Đại học Cần Thơ

dtahuy@ctec.edu.vn, clsalin@ctec.edu.vn, nhhoa@ctu.edu.vn, ntnghe@cit.ctu.edu.vn

TÓM TẮT *— Hệ thống gợi ý (Recommender Systems – RS) được ứng dụng khá thành công trong thương mại điện tử, nó đưa ra dự đoán về các mục thông tin (items) mà người dùng có thể thích theo một trong hai cách – phản hồi tường minh và phản hồi tiềm ẩn. Phản hồi tường minh dựa vào các đánh giá, xếp hạng,... của người dùng trong quá khứ lên các items để gợi ý các items mà người dùng đó có thể thích trong tương lai. Phản hồi tiềm ẩn dựa vào các items mà người dùng từng lựa chọn, tham khảo hay xem các items đó để đưa ra các gợi ý cho người dùng. Tuy nhiên, vấn đề khó khăn chung của hầu hết hệ thống gợi ý là khi người dùng mới chưa có bất kỳ phản hồi nào trong hệ thống thì hầu như hệ thống không đưa ra gợi ý chính xác cho họ, đó chính là vấn đề khởi đầu lạnh hay còn gọi là vấn đề "Cold-start". Trong bài viết này chúng tôi giới thiệu một giải pháp trong việc xử lý vấn đề người dùng mới dựa trên các thông tin, các thuộc tính (attributes) của người dùng mới. Qua đó, chúng tôi xây dựng các thực nghiệm để kiểm chứng tính khả thi của các mô hình. Kết quả thực nghiệm cho thấy giải pháp đề xuất có khả năng gợi ý khá tốt cho những người dùng mới.*

Từ khoá *— Hệ thống gợi ý, khởi đầu lạnh, vấn đề người dùng mới.*

I. GIỚI THIỆU

Việc ứng dụng hệ thống gợi ý (recommender systems - RS) trong thương mại điện tử đã mang lại sự thành công và nhiều sự lựa chọn cho người dùng. Ví dụ trong hệ thống bán hàng trực tuyến, với hàng ngàn sản phẩm khác nhau, đa dạng về mẫu mã, chất lượng, tiêu chí,…để khách hàng mua được một sản phẩm ưng ý thì một lời tư vấn sẽ là rất quan trọng, hệ thống gợi ý đóng vai trò như một người trung gian sẽ đưa ra các tư vấn cho khách hàng. Thực chất của hệ thống này là hỗ trợ người dùng đưa ra quyết định dựa trên những thông tin mà nó thu thập được từ những "hành vi" hoặc dựa trên những phản hồi của người dùng trong quá khứ. Những phản hồi của người dùng có thể là tường minh (explicit) hay tìm ẩn (implicit). Phản hồi tường minh do người dùng trực tiếp đánh giá lên item như: bình chọn, xếp hạng cho các items yêu thích…; phản hồi tiềm ẩn do hệ thống tự động thu thập dựa trên hành vi của người dùng chẳng hạn như số lần click chuột vào một item nào đó, thời gian xem xét một item, số lần mua các items…

Có rất nhiều giải thuật xây dựng hệ thống gợi ý, tuy nhiên có thể phân thành ba nhóm chính [6][8]: *Một là*, gợi ý dựa trên nội dung: Hệ thống gợi ý dựa vào nội dung đưa ra những kết quả là những sản phẩm có nhiều điểm tương tự với những sản phẩm mà người dùng đã lựa chọn trong quá khứ. *Hai là*, gợi ý dựa trên cộng tác: hệ thống gợi ý lọc cộng tác khai thác thông tin về những hành động trong quá khứ hoặc ý kiến của cộng đồng người dùng để đưa ra dự đoán những sản phẩm mà người dùng hiện tại có thể thích. *Ba là*, gợi ý dựa trên cách tiếp cận kết hợp nhiều mô hình dự đoán trong hệ thống gợi ý.

Tuy nhiên, khó khăn lớn của những phương pháp gợi ý dựa trên lọc cộng tác đang gặp phải là khi người dùng mới chưa có bất kỳ phản hồi, đánh giá nào trong hệ thống, vì vậy hệ thống không có dữ liệu huấn luyện và không thể đưa ra dự đoán cho người dùng mới, đó chính là vấn đề khởi đầu lạnh (cold-start) hay còn gọi là vấn đề "Người dùng mới" (new user) hay "Mục tin mới" (new item) trong hệ thống gợi ý. Để khắc phục những khó khăn trong vấn đề này, đã có nhiều nghiên cứu đưa ra các giải pháp xử lý khác nhau sẽ được chúng tôi trình bày chi tiết trong mục III.

Trong bài viết này, chúng tôi đề xuất giải pháp xử lý vấn đề người dùng mới, sử dụng kỹ thuật phân rã ma trận (matrix factorization) – một kỹ thuật thành công nhất (state-of-the-art) trong hệ thống gợi ý [13] - kết hợp với các thuộc tính (attributtes) của người dùng để tìm độ tương đồng của người dùng mới với các người dùng khác trong hệ thống hoặc kết hợp với kỹ thuật hồi quy tuyến tính, từ đó đưa ra gợi ý cho họ. Các tập dữ liệu chuẩn được sử dụng để đánh giá phương pháp đã đề xuất và so sánh với các phương pháp thường được dùng nhất trong xử lý người dùng mới là phương pháp Global Average và Item Average.

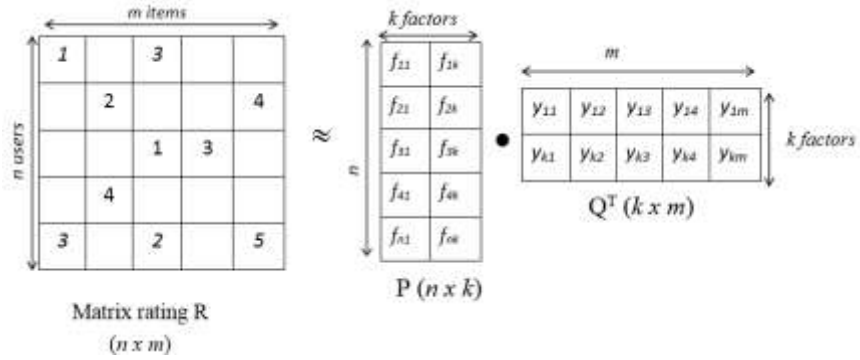
II. HỆ THỐNG GỢI Ý (RECOMENDER SYSTEMS)

Hệ thống gợi ý (Recommender Systems - RS) là một dạng của hệ thống lọc thông tin (information filtering), nó được sử dụng để dự đoán sở thích (preferences) hay xếp hạng (rating) mà người dùng (user) có thể dành cho một mục thông tin (item) nào đó mà họ chưa xem xét tới trong quá khứ (item có thể là bài báo, bộ phim, đoạn video clip, sách,..) [15] nhằm đưa ra những gợi ý về những mục tin phù hợp cho người dùng.

Thông tin về user-item-rating thường được biểu diễn thông qua một ma trận (rating matrix) mà ở đó mỗi dòng là một user, mỗi cột là một item, mỗi ô là một giá trị đánh giá đại diện cho mức độ "thích" của user dành cho item tương ứng. Các ô trống là người dùng chưa có đánh giá trên item đó. Thông thường các ô có giá trị rất ít, vì người dùng rất ít đánh giá trong quá khứ, do vậy tạo nên một ma trận cực thưa (sparse matrix), như minh họa bên trái của Hình 1. Gọi

$R^{|U|x\ |I|}$ là ma trận đánh giá của các users đối với các items, r_{ij} là đánh giá của user i đối với item j nào đó. Ví dụ: với i=2, j=4, r_{24} = 3, nghĩa là người dùng thứ 2 đã đánh giá sản phẩm thứ 4 với mức đánh giá là 3. Mức độ đánh giá thường tỷ lệ thuận với mức độ yêu thích của người dùng.

Một trong những kỹ thuật hay được dùng trong RS là lọc cộng tác (collaborative filtering). Phương pháp này dựa trên thông tin tương tác (như mua, bán, hoặc đánh giá,..) của người dùng để đưa ra các dự đoán dựa trên độ tương đồng về sở thích giữa người dùng với nhau, trong đó thành công nhất là kỹ thuật phân rã ma trận (Matrix Factorization - MF). MF là việc chia một ma trận lớn R thành hai ma trận có kích thước nhỏ hơn P và Q, sao cho ta có thể xây dựng lại R từ hai ma trận nhỏ hơn này càng chính xác càng tốt, nghĩa là $R \sim P.Q^T$ [13]. Kỹ thuật MF được mô tả như trong hình 1.



Hình 1. Kỹ thuật phân rã ma trận

MF phân rã ma trận rating $R^{|U|x\ |I|}$ thành 2 ma trận $P^{|U|\ x\ K}$ và $Q^{|I|\ x\ K}$, trong đó P là một ma trận mà mỗi dòng u là một vector gồm K nhân tố tiềm ẩn (latent factors) mô tả người dùng u, Q là một ma trận mà mỗi dòng i là một vector gồm K nhân tố tiềm ẩn mô tả mục thông tin i; Khi đó, để tính rating của ma trận R ta áp dụng công thức sau:

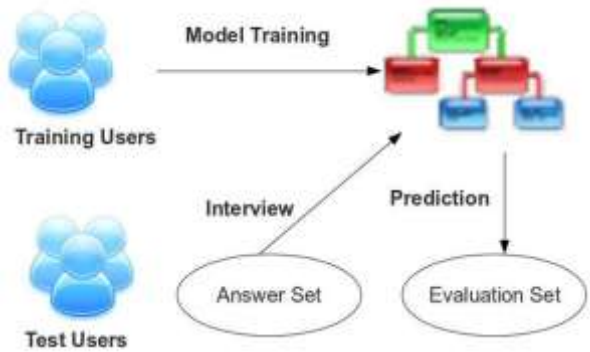
$$r_{ui} = \sum_{k=1}^{K} p_{uk} \cdot q_{ik} \quad (1)$$

trong đó: $\hat{r}_{ui}$ là giá trị dự đoán của người dùng u đối với item i; p, q: các vector trong ma trận P và Q.

Mặc dù kỹ thuật này đã được áp dụng rất thành công, tuy nhiên nó vẫn còn hạn chế như những kỹ thuật lọc cộng tác khác, đó là vấn đề người dùng mới (hay cũng là vấn đề khởi đầu lạnh "cold–start"). Vấn đề này xảy ra khi một người dùng mới hoặc một sản phẩm mới tham gia vào hệ thống; do không có đủ thông tin nên sẽ rất khó khăn để xác định những người dùng (hoặc sản phẩm) tương đồng với người dùng mới (hoặc sản phẩm mới) này. Thông thường trong MF, khi gặp phải vấn đề này, người ta dùng các kỹ thuật như Global Average, Item Average,.. để dự đoán kết quả cho người dùng mới. Một số phương pháp phức tạp khác cũng đã được đề xuất như trình bày dưới đây.

III. NHỮNG NGHIÊN CỨU LIÊN QUAN

Một trong những nghiên cứu giải quyết vấn đề Cold-start là kỹ thuật FMF (Fucntion Matrix Factorization) [21] sử dụng mô hình cây quyết định, tại mỗi nút trong cây sẽ đưa ra các truy vấn và người dùng sẽ chọn câu trả lời để đến các nút con tiếp theo, từ đó hệ thống đưa ra gợi ý cho người dùng.



Hình 2. Mô hình giải quyết vấn đề cold-start user của kỹ thuật FMF

Nghiên cứu của Zeno Gantner, et.al [9] sử dụng "Bayesian Personalized Ranking" để huấn luyện mô hình dự đoán dựa vào ánh xạ từ các thuộc tính của các đối tượng trong hệ thống đến các nhân tố tiềm ẩn của ma trận để tìm ra đánh giá cho đối tượng mới. Một phương pháp khác là Quickstep, là một hệ thống gợi ý lai (hybrid recommender system), giải quyết các vấn đề thực tế của việc gợi ý bài báo khoa học trực tuyến để các nhà nghiên cứu có thể tìm thấy chúng. Hành

động duyệt web của người dùng sẽ được âm thầm giám sát thông qua một máy chủ proxy ở mỗi đăng nhập URL thông qua trình duyệt trong các hoạt động bình thường. Các phản hồi tường minh và URL đã duyệt là thông tin cơ bản lưu sở thích của mỗi người dùng. Nghiên cứu [11] xây dựng hệ thống gợi ý tích hợp Quickstep [15], là sự kết hợp của AKT ontology và hệ thống OntoCoPI đã chứng minh một cách tiếp cận mới của các tác giả để giảm cold-start của hệ thống. Một kỹ thuật khác cũng giải quyết khá tốt vấn đề cold-start là Context-aware Semi-supervised Co-training algorithm (CSEL) [20], kỹ thuật này xây dựng mô hình có khả năng tăng độ chính xác của các dự đoán bằng cách kết hợp ngữ cảnh với kỹ thuật đồng huấn luyện có giám sát và bán giám sát. Nghiên cứu [19] đề xuất mô hình hồi quy dựa vào thông tin của các cặp giữa user/item. Các thông tin của users như tuổi, giới tính, nghề nghiệp… và các thông tin của items như tên sản phẩm, nhà sản xuất, năm sản xuất,…sau đó đưa ra tiên đoán để giải quyết vấn đề cold-start.

Martin Saveski và Amin Mantrach trong nghiên cứu [14] đã đưa ra một phương pháp giải quyết vấn đề Cold-start đối với items. Tác giả dựa vào những đặc tính của sản phẩm và sự lựa chọn, đánh giá của người dùng trong hệ thống đối với những items có đặc tính tương tự với items mới để huấn luyện cho mô hình, từ đó đưa ra gợi ý cho những items mới. Một phương pháp khác cũng giải quyết vấn đề cold-start trong hệ thống gợi ý đó được đề cập trong nghiên cứu [7] bằng cách sử dụng mô hình mạng xã hội (Network Sub-community) kết hợp với luật quyết định (decision ontology). Mạng xã hội thực hiện nhiệm vụ phân tích thông tin của các người dùng trong hệ thống và tìm ra mối tương quan giữa các người dùng với nhau. Theo đó, kiến trúc miền quyết định xây dựng mô hình cơ bản dựa trên thông tin của những người dùng đã có từ đó đưa ra gợi ý cho người dùng mới.

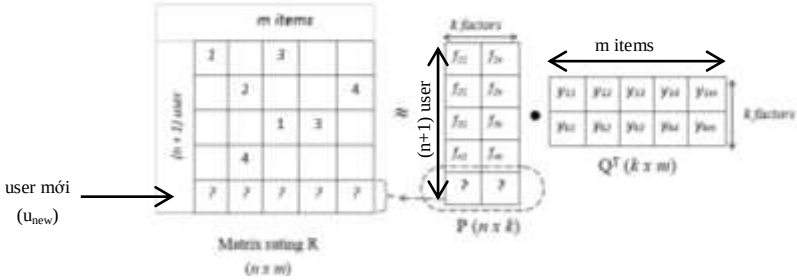
Nhìn chung có nhiều phương pháp giải quyết vấn đề cold-start, tuy nhiên các phương pháp thường kết hợp với các mô hình xử lý khá phức tạp, điển hình là trong các mô hình quyết định [21], hệ thống gợi ý phải tương tác với người dùng để đánh giá các câu trả lời, hay trong nghiên cứu [20] sử dụng kỹ thuật đồng huấn luyện bán giám sát và có giám sát cũng áp dụng mô hình dựa vào ngữ cảnh từ đó mới đưa ra được dự đoán. Còn trong nghiên cứu [7] thì phải kết hợp với mô hình mạng xã hội khá phức tạp để tìm ra mối tương quan giữa các người dùng, từ đó mới đưa ra được gợi ý cho họ.

Trong nghiên cứu này, chúng tôi đề xuất một hướng tiếp cận khác để xử lý vấn đề người dùng mới như trình bày dưới đây.

IV. GIẢI PHÁP ĐỀ XUẤT

Chúng tôi đề xuất hai mô hình để xử lý vấn đề cold-start đối với người dùng mới (new user), trường hợp với các mục thông tin mới (new item) cũng có thể được áp dụng tương tự. Giải pháp này là sự phối hợp giữa kỹ thuật phân rã ma trận (matrix factorixation) với việc dùng láng giềng lân cận (k-nearest neighbors) hay hồi quy tuyến tính (linear regression) để dự đoán các nhân tố của người dùng mới thông qua các thuộc tính của họ.

Như minh họa trong hình 3 dưới đây, với người dùng mới (người thứ n+1) thì kỹ thuật MF hoàn toàn không thể xác định được nhân tố của người dùng này (user factor - dòng cuối trong ma trận P), vì vậy không thể đưa ra dự đoán cho họ.

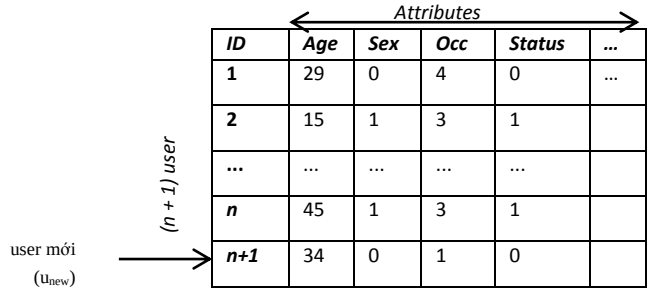


Hình 3. Minh họa kỹ thuật phân rã ma trận cho người dùng mới

Để giải quyết vấn đề này, dựa trên ý tưởng từ bài viết [9], thông qua các thuộc tính của người dùng mới chúng tôi giới thiệu 2 cách:

- Tìm các người dùng tương đồng với người dùng mới (dùng phương pháp kNN trên thuộc tính của họ) sau đó sử dụng các nhân tố (user factors) của những người dùng tương đồng để tìm nhân tố cho người dùng mới. Phương pháp này được đặt tên là MF-kNN.
- Sử dụng các thuộc tính của người dùng để xây dựng mô hình hồi quy, từ đó dự đoán từng nhân tố (factor) cho người dùng mới. Phương pháp này được đặt tên là MF-LR.

Chi tiết của từng phương pháp sẽ được trình bày dưới đây. Trước hết chúng tôi minh họa ma trận thuộc tính của user có dạng như trong Hình 4:



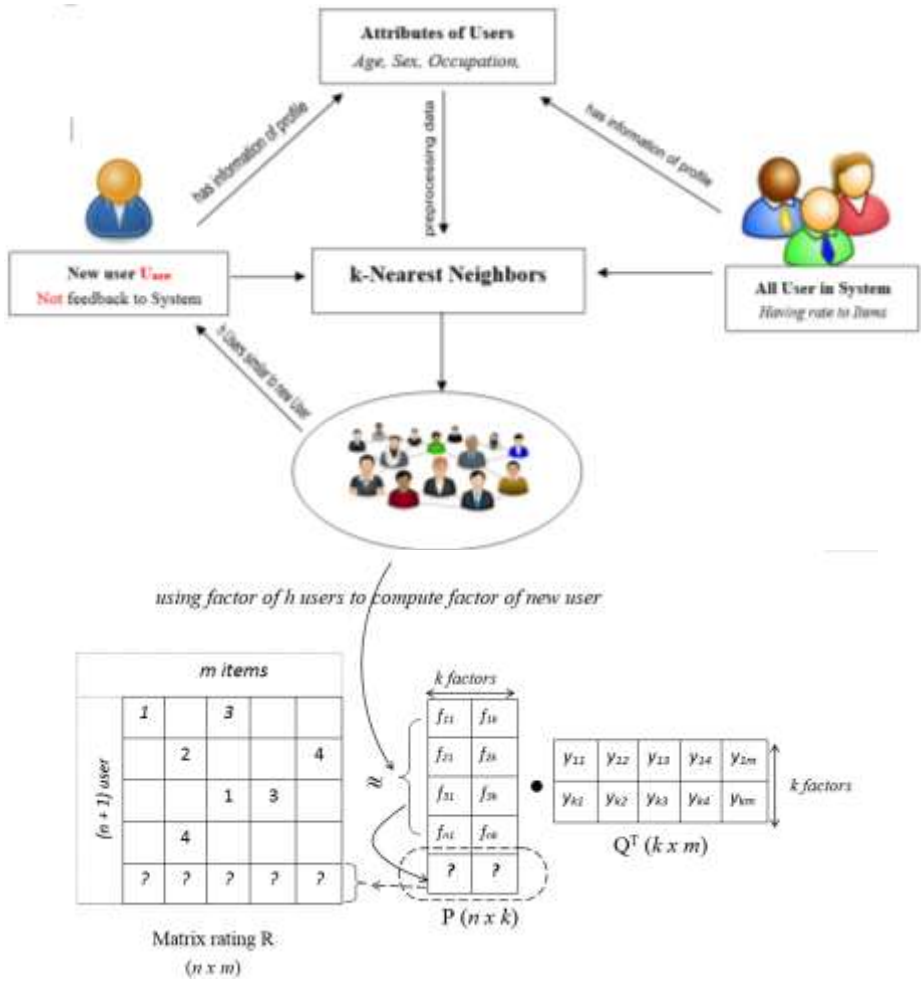
ID	Age	Sex	Occ	Status	…
1	29	0	4	0	…
2	15	1	3	1	
…	…	…	…	…	
n	45	1	3	1	
n+1	34	0	1	0	

Hình 4. Ma trận thuộc tính của user

Gọi $A^{|U| \times T}$ là ma trận lưu trữ thuộc tính của các user, trong đó giá trị các thuộc tính của user được chuẩn hóa để đưa các thuộc tính về dạng các trị số có thể dễ dàng ước lượng, tính toán. Ngoài ra, việc chuẩn hóa này còn giúp giá trị các thuộc tính được chuyển vào một miền giá trị xác định cho trước, giúp dữ liệu được cân bằng hơn. Ví dụ khi so sánh hai người dùng u_1 (độ tuổi là 50, giới tính nam, nhóm nghề 2, độc thân) với người dùng u_2 (20 tuổi, giới tính nữ, nhóm nghề 1, độc thân) thì rõ ràng thuộc tính độ tuổi sẽ chiếm ưu thế hơn các giá trị của các thuộc tính khác và do đó khi chúng ta áp dụng các công thức tìm độ tương đồng của các người dùng dựa vào giá trị các thuộc tính thì thuộc tính độ tuổi (age) sẽ ảnh hưởng nhiều nhất đến độ tương đồng của các người dùng.

*** Phương pháp 1: Dự đoán các nhân tố tiềm ẩn (latent factors) bằng phương pháp láng giềng lân cận "k-Nearest Neighbors" (MF-kNN)**

Mô hình chi tiết được trình bày trong sơ đồ sau đây:



Hình 5. Mô hình dự đoán nhân tố tiềm ẩn bằng phương pháp MF-kNN

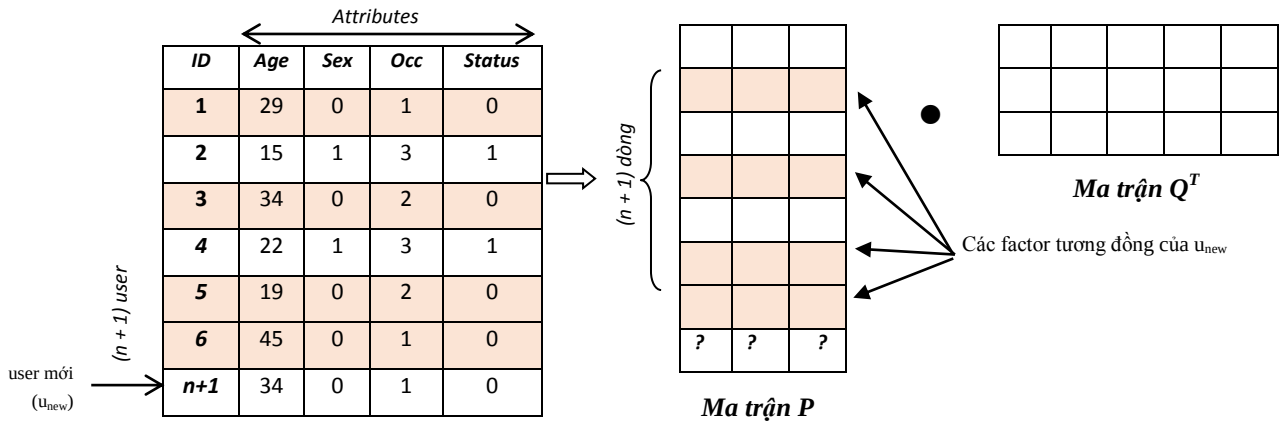
Ý tưởng của phương pháp là phân loại một người dùng vào trong lớp tương đồng với nó nhất dựa vào độ tương đồng các thuộc tính của người dùng [25]. Trong trường hợp u_{new} là người dùng mới xuất hiện trong hệ thống chưa có bất cứ đánh giá cho các items nào, chúng tôi đề xuất dựa vào giá trị các thuộc tính (attributes) của user như: tuổi, giới tính, nghề nghiệp,... để tìm độ tương đồng giữa user mới (u_{new}) với tất cả các user trong tập dữ liệu, sau khi xác định được những user tương đồng nhất, ta sử dụng nhân tố của những user tương đồng đó để xác định nhân tố cho user mới.

Trong bài viết này chúng tôi dùng độ đo cosine (cosine similarity) để xác định độ tương đồng của hai người dùng u_a và u_b như công thức:

$$sim(u_a, u_b) = cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2 \|\vec{b}\|_2} = \frac{\sum_i^m p_{ai} \cdot p_{bi}}{\sqrt{\sum_i^m p_{ai}^2} \sqrt{\sum_i^m p_{bi}^2}} \quad (3)$$

trong đó: p_{ai}, p_{bi} là giá trị thuộc tính thứ i của người dùng a và b.
m là số thuộc tính của các vector.

Giả sử sau khi dựa vào giá trị các thuộc tính của user để tìm độ tương đồng, với K=4 ta được tập các user tương đồng với u_{new} là: $H\{u_1, u_3, u_5, u_6\}$ => Trong ma trận phân rã P (với R~P.Q^T), các dòng tương ứng với các UID tương đồng với u_{new} cũng là tập hợp các factor tương đồng của u_{new}.



Hình 6. Xác định factor tương đồng của người dùng mới dựa vào độ tương đồng của thuộc tính

Dựa vào độ tương đồng của u_{new} và các user khác, chúng ta tìm được tập H các user tương đồng với u_{new}. Khi đó chúng ta có thể tính được factor của u_{new} trong ma trận P dựa vào công thức (4);

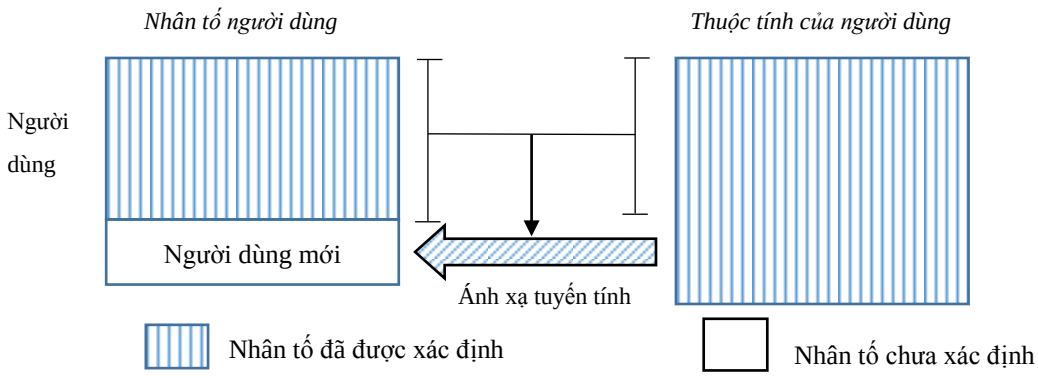
$$f_{u_{new}k} = \frac{\sum_{u' \in H_{u_{new}}} sim(u_{new}, u') . f_{u'k}}{\sum_{u' \in H_{u_{new}}} |sim(u_{new}, u')|} \quad (4)$$

trong đó: $f_{u_{new}k}$ *: factor dự đoán của* u_{new} *đối với nhân tố tiềm ẩn thứ k.*
$sim(u_{new}, u')$ *: độ tương đồng của người dùng* u_{new} *và u'.*
$f_{u'k}$*: factor của người dùng u' đối với nhân tố tiềm ẩn thứ k.*
H: tập các user tương đồng với u_{new}

Sau khi tìm được giá trị các factor của u_{new} chúng ta dễ dàng tính lại được rating của u_{new} theo công thức (1)

*** Phương pháp 2: Dự đoán các nhân tố bằng mô hình hồi quy tuyến tính (MF-LR)**

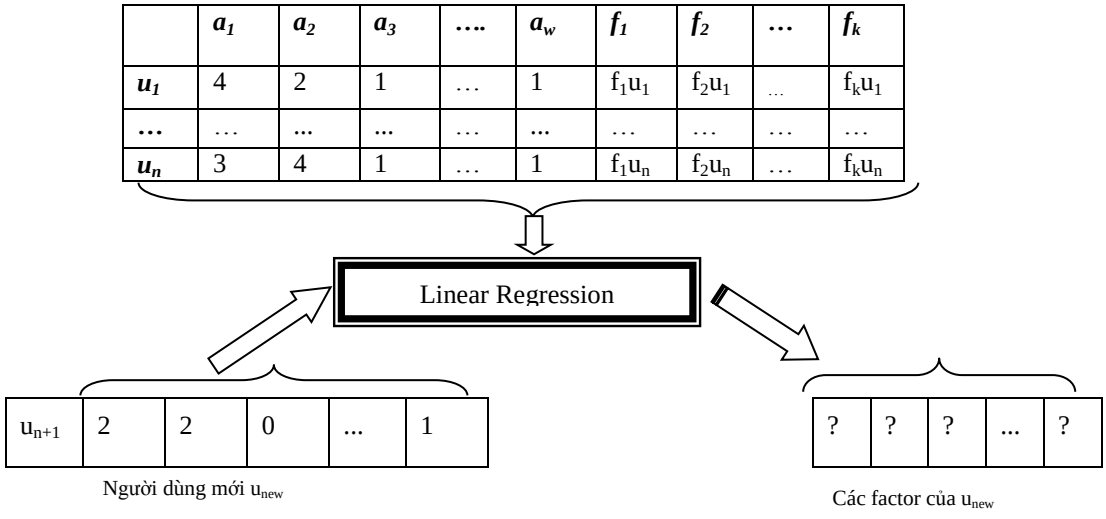
Ý tưởng của phương pháp này là dựa vào giá trị các thuộc tính và các nhân tố (factors) của các user có trong hệ thống để tính toán sự ảnh hưởng tuyến tính đến các nhân tố của người dùng mới.



Hình 7. Phương pháp mapping tuyến tính từ thuộc tính sang nhân tố của người dùng

Trong mô hình này, chúng tôi sử dụng các thuộc tính và các nhân tố đã xác định của những người dùng trong hệ thống kết hợp với kỹ thuật hồi quy tuyến tính để ánh xạ đến các factor của người dùng mới. Gọi A = { a_1, a_2, a_3,…, a_w} là tập hợp tất cả thuộc tính của user (a_i là thuộc tính thứ i); P và Q là các ma trận được phân rã bằng kỹ thuật phân rã ma trận từ ma trận rating R, với R~ $P.Q^T$.

Sau khi chuẩn hóa giá trị các thuộc tính về miền giá trị xác định cho trước, ta xây dựng mô hình dự đoán nhân tố của người dùng mới minh họa trong hình sau:



	a_1	a_2	a_3	….	a_w	f_1	f_2	…	f_k
u_1	4	2	1	…	1	f_1u_1	f_2u_1	…	f_ku_1
…	…	…	…	…	…	…	…	…	…
u_n	3	4	1	…	1	f_1u_n	f_2u_n	…	f_ku_n

u_{n+1}	2	2	0	...	1

?	?	?	...	?

Hình 8. Dự đoán các factor của người dùng mới dựa trên factors của các người dùng đã có

Trong mô hình này, chúng tôi dựa vào giá trị các thuộc tính (a_1, a_2,..., a_w) (xem như các predictors) kết hợp với các user factor (f_1, f_2,...,f_k) (xem như các thuộc tính đích – target class/attribute) được tạo ra từ kỹ thuật phân rã ma trận, sau đó dùng Linear regression để xác định các factor của người dùng mới. Sau khi tìm được các giá trị các factor của u_{new}, ta hoàn toàn có thể dự đoán được xếp hạng của người dùng mới này theo công thức (1) của kỹ thuật phân rã ma trận thông thường.

V. KẾT QUẢ THỰC NGHIỆM

a. Dữ liệu dùng để đánh giá

- Tập dữ liệu của hệ thống gợi ý phim Movielens 100k và 1M (grouplens.org/datasets/movielens): Tập dữ liệu 100K có 100.000 đánh giá được thực hiện bởi 943 người dùng trên số lượng 1.682 phim; Tập dữ liệu 1M có 1.000.209 đánh giá được thực hiện bởi 6.040 người dùng trên số lượng 2000 bộ phim, mỗi người dùng có đánh giá ít nhất 20 phim và mức đánh giá từ 1..5;

- Tập dữ liệu Restaurant & Consumer data (RCData, archive.ics.uci.edu/ml/machine-learning-databases/00232): Tập dữ liệu này được thu thập từ một phần của hệ thống gợi ý nhà hàng theo đánh giá của khách hàng tại các thành phố trên đất nước México, mục tiêu tạo ra một danh sách Ntop nhà hàng tốt nhất theo bình chọn của khách hàng. Thông tin các tập dữ liệu dùng đê thực nghiệm như trình bày trong bảng sau

Bảng 1. Mô tả tập dữ liệu thực nghiệm

Data Set	Users	User-Attributes	Items	Items-Attributes	Rating
MovieLens 100K	943	3	1682	19	100,000 [1-5]
MovieLens 1M	6040	3	2000	19	1,000,209 [1-5]
RC Data	138	19	130	21	1,161 [1-3]

Các tập dữ liệu nên được chuẩn hóa trước khi đưa vào mô hình dự đoán.

b. Phương pháp chuẩn hóa dữ liệu đầu vào

Có nhiều phương pháp chuẩn hóa dữ liệu như z-score normalization, normalization by decimal scaling, hay min-max normalization..., trong bài viết này chúng tôi thực hiện chuẩn hóa dữ liệu theo công thức MIN-MAX normalize như sau:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

trong đó:

v: giá trị cũ, v ∈[minA, maxA]; v': giá trị mới, v' ∈ [new_min_A, new_maxA];

Việc chuẩn hóa này giúp dữ liệu đầu vào đúng định dạng và giá trị của nó thuộc miền giá trị xác định nhằm tránh độ lệch (biases) khi dùng các độ đo tương đồng (như trong kNN).

c. Các độ đo dùng trong thực nghiệm

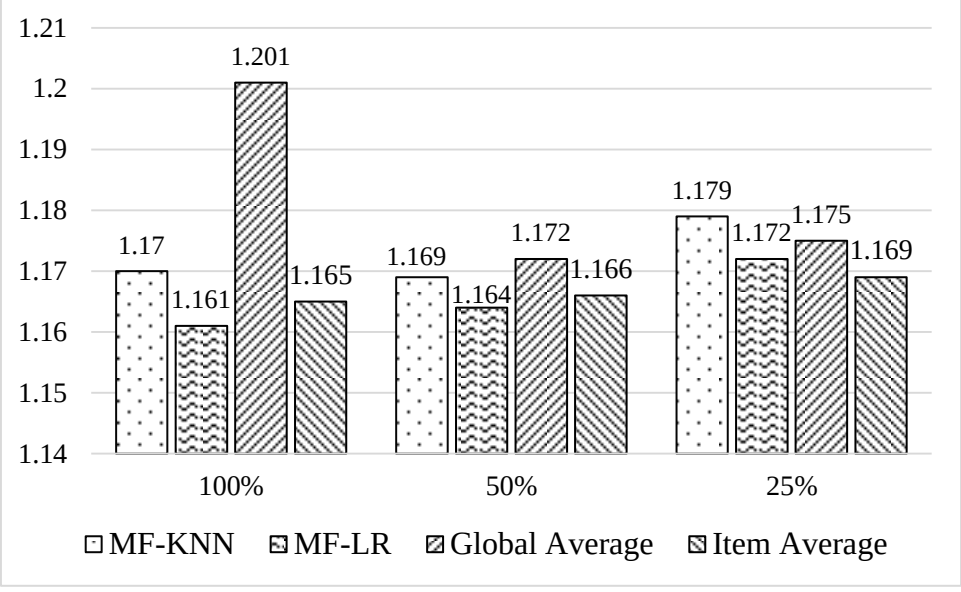
Có nhiều phương pháp khác nhau mà chúng ta có thể sử dụng để đo độ lỗi của giải thuật như: Root Mean Squared Error (RMSE), MAE (Mean Absolute Error), F-Measure, Area Under the ROC curve (AUC),… nhưng mỗi phương pháp đánh giá sẽ thích hợp cho từng lĩnh vực khác nhau. Ở đây sử dụng độ đo phổ biến là RMSE và MAE để đánh giá hiệu quả của các phương pháp.

d. Phương pháp đánh giá

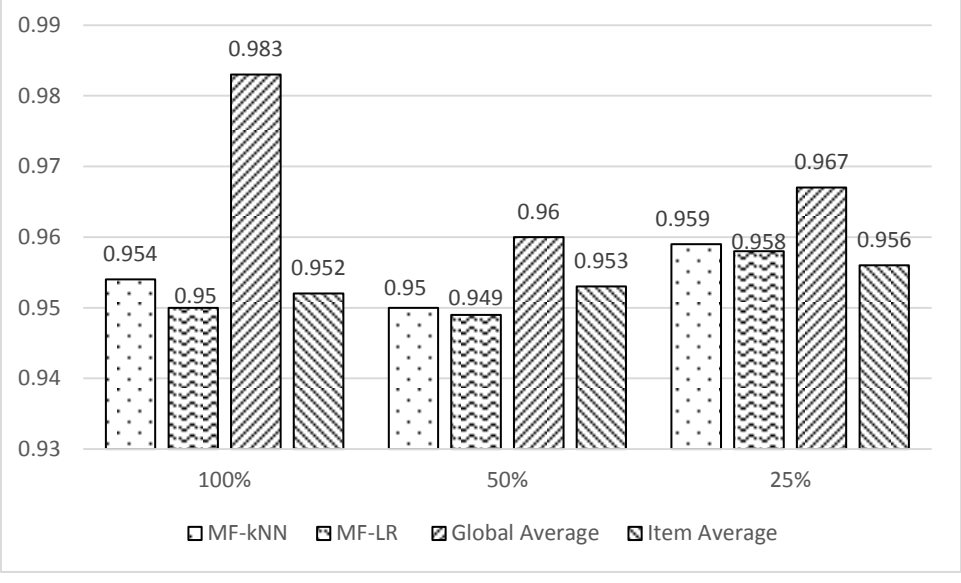
Chúng tôi sử dụng kỹ thuật đánh giá phổ biến là nghi thức k-folds cross validation. Trong quá trình kiểm tra, để kiểm chứng hiệu quả của mô hình giải quyết vấn đề Cold-start chúng tôi phân dữ liệu trong tập test lần lượt thành các bộ dữ liệu có chứa: 25%, 50%, và 100% người dùng mới; Kết quả đánh giá được so sánh với các phương pháp Global Average và Item Average là những phương pháp đơn giản thường được dùng trong trường hợp đánh giá cho người dùng mới và mục tin mới.

e. Kết quả thực nghiệm

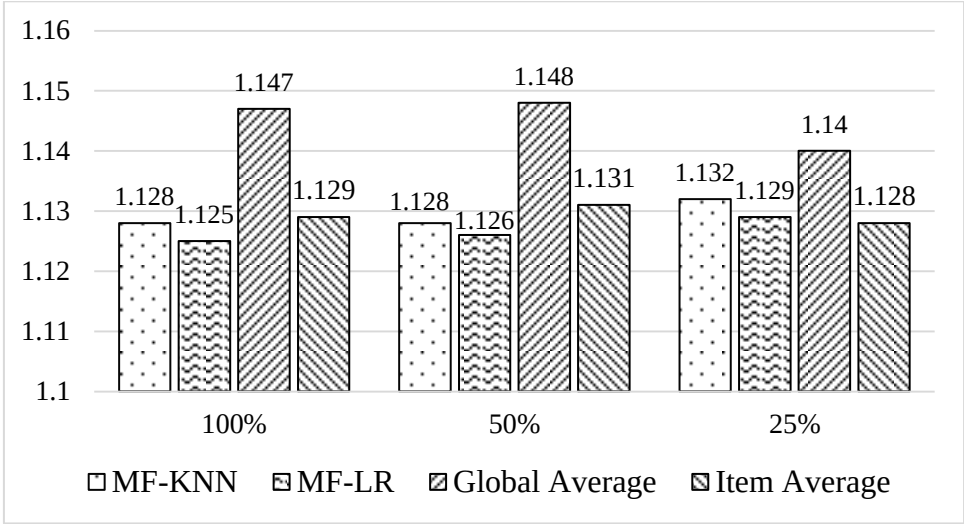
Kết quả thực nghiệm được trình bày như trong các hình sau:



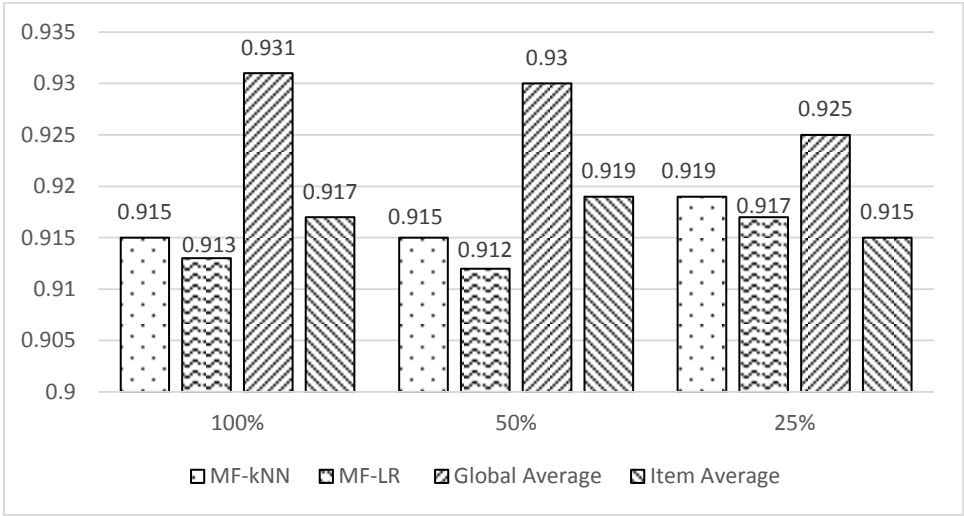
Hình 9. Thể hiện độ đo lỗi RMSE trong tập dữ liệu MovieLens 100K



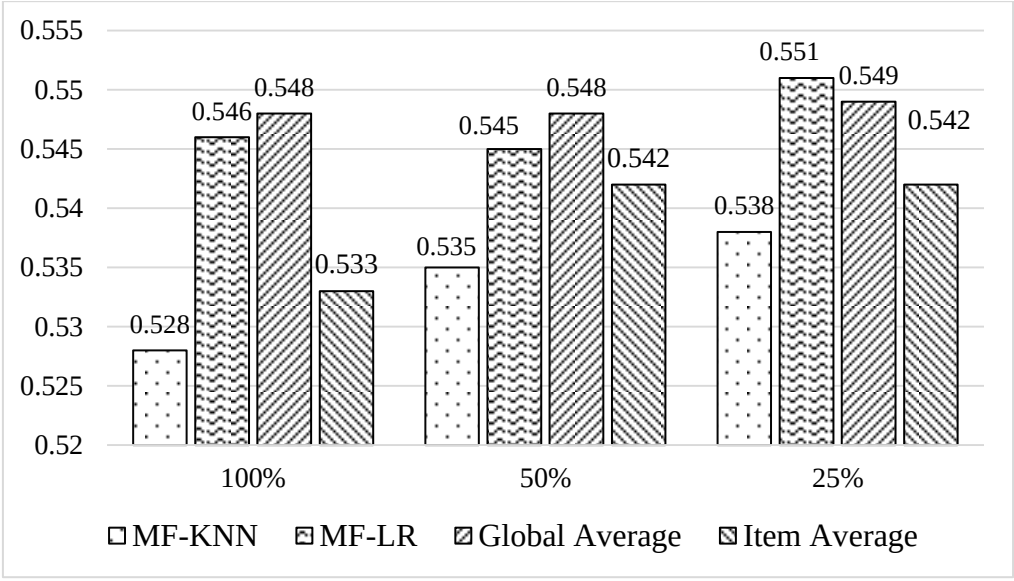
Hình 10. Thể hiện độ đo lỗi MAE trong tập dữ liệu MovieLens 100K



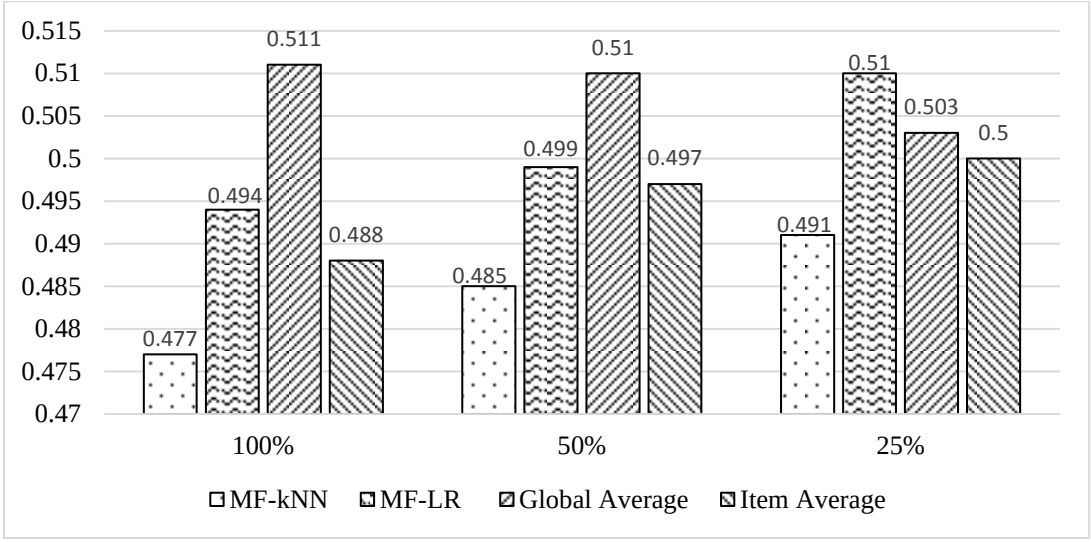
Hình 11. Thể hiện độ đo lỗi RMSE trong tập dữ liệu Movie Lens 1M



Hình 12. Thể hiện độ đo lỗi MAE trong tập dữ liệu Movie Lens 1M



Hình 13. Thể hiện độ đo lỗi RMSE trong tập dữ liệu RC Data



Hình 14. Thể hiện độ đo lỗi MAE trong tập dữ liệu RC Data

Kết quả thực nghiệm trên đã cho thấy các phương pháp đề xuất có khả năng làm giảm lỗi cho mô hình dự đoán ***trong xử lý vấn đề Cold-start trong hệ thống gợi ý***. Cụ thể, trên tập dữ liệu MovieLens 100K và MovieLens 1M, đối với tỷ lệ 100% người dùng mới thì phương pháp MF-LR cho kết quả tốt hơn các phương pháp còn lại, đối với tỷ lệ 50% và 25% người dùng mới thì cả 2 phương pháp đề xuất đều cho kết quả tốt hơn so với Global Average. Trên tập dữ liệu RC Data, tương ứng với các tập dữ liệu có tỷ lệ người dùng mới là 100%, 50% và 25% thì phương pháp MF-KNN đều cho kết quả tốt hơn các phương pháp còn lại.

Thật vậy, do tập dữ liệu MovieLens nhiều hơn trong tập dữ liệu RC Data nên khi dùng MF-LR dựa trên mô hình hồi quy tuyến tính sẽ cho kết quả dự đoán tốt hơn, ngược lại trong tập dữ liệu RC Data các user có nhiều thuộc tính hơn tập MovieLens nên khi áp dụng phương pháp MF-KNN để tìm các user tương đồng với người dùng mới dựa vào thuộc tính sẽ cho kết quả tốt hơn.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Qua các mô hình mà chúng tôi đề xuất để xử lý vấn đề Cold-start trong hệ thống gợi ý đã phần nào giải quyết được các hạn chế trong việc thiếu thông tin đánh giá của người dùng trong hệ thống bằng cách dựa vào thông tin cá nhân hay thuộc tính của đối tượng. Từ đó có thể sử dụng kỹ thuật láng giềng lân cận hoặc hồi quy tuyến tính, kết hợp với kỹ thuật phân rã ma trận để đưa ra dự đoán cho người dùng. Từ các kết quả thực nghiệm cho thấy mô hình đề xuất có độ lỗi thấp hơn mô hình gợi ý truyền thống (baseline) và không phụ thuộc nhiều vào tỷ lệ người dùng mới trong hệ thống. Nhìn chung các mô hình đều giải quyết được vấn đề Cold-start trong RS, tuy nhiên đối với các tập dữ liệu mà đối tượng có nhiều thuộc tính sẽ cho kết quả chính xác hơn. Việc áp dụng các mô hình đề xuất vào các hệ thống gợi ý hoàn toàn khả thi để xử lý cho các trường hợp cold-start. Đối với các item mới trong hệ thống, chúng tôi đề xuất phương pháp xử lý tương tự.

Từ kết quả thực nghiệm trên, để xử lý tốt hơn cho các trường hợp cold-start trong hệ thống gợi ý, hướng nghiên cứu tiếp theo của đề tài là phân tích sâu hơn các phương pháp giải quyết vấn đề cold-start của các tác giả khác để làm cơ sở đối chiếu với kết quả của mô hình đã đề xuất, từ đó chọn lọc, xử lý thêm nhiều thuộc tính có liên quan đến đối tượng cần gợi ý và có thể kết hợp với các kỹ thuật gợi ý khác để giúp hệ thống đưa ra dự đoán chính xác hơn. Bên cạnh đó, cần kiểm chứng mô hình trên nhiều tập dữ liệu khác nhau và so sánh kỹ thuật đã đề xuất với các phương pháp giải quyết vấn đề cold-start khác.

TÀI LIỆU THAM KHẢO

[1]. G. Adomavicius and 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,pp.734–749, 2005.

[2]. Asanov, Daniar. "Algorithms and methods in recommender systems." Berlin Institute of Technology, Berlin, Germany (2011).

[3]. Andrew I. Schein, Alexandrin Popescul, and Lyle H. Ungar, David M. Pennock. “Methods and Metrics for Cold-Start Recommendations”. In SIGIR, pages 253-260. ACM, 2002.

[4]. D. Agarwal and B.-C. Chen. “Regression based latent factor models”. ACM KDD 2009

[5]. D. Billsus and M. Pazzani, “Learning collaborative information filters,” in Proceedings of the 15th International Conference on Machine Learning (ICML '98), 1998.

[6]. Li Chen, Guanliang Chen, and Feng Wang. “Recommender systems based on user reviews: the state of the art”. User Modeling and User-Adapted Interaction 25, pages: 99-154, 2015.

[7]. Meng, Chen, et al. "A Method to Solve Cold-Start Problem in Recommendation System based on Social Network Sub-community and Ontology Decision Model." 3rd International Conference on Multimedia Technology (ICMT-13). Atlantis Press, 2013.

[8]. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Recommender Systems Handbook (1st ed.). Springer-Verlag New York, Inc., New York, NY,USA, 2010

[9]. Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle and Lars Schmidt-Thieme. "Learning Attribute-to-Feature Mappings for Cold-Start Recommendations", pages 176 – 185, IEEE, 2012.

[10]. Shien Ge and Xinyang Ge. "An SVD-based Collaborative Filtering approach to alleviate cold-start problems", 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pages: 1474 – 1477, IEEE, 2012.

[11]. Guarino, N., Masolo, C. and Vetere, G. OntoSeek: "Content-Based Access to the Web", IEEE Intelligent Systems, Vol. 14,No. 3, May/June 1999.

[12]. Koren. Y. "Factor in the neighbors: Scalable and accurate collaborative filtering", AT&T Labs - Research 180 Park Ave, Florham Park, NJ 07932, 2010

[13]. Koren. Y, Bell. R, 2009: "Matrix Factorization Techniques for Recommender Systems", pages 42 – 49, IEEE.

[14]. Martin Saveski, Amin Mantrach: "Item cold-start recommendations: learning local collective embeddings". RecSys '14 Proceedings of the 8th ACM Conference on Recommender systems. Pages 89-96. ACM, 2014.

[15]. Nguyen Thai-Nghe. An introduction to factorization technique for building recommendation systems. Vol. 6/2013, pages: 44-53, Journal of Science - University of Da Lat, ISSN 0866-787X, 2013

[16]. Nguyễn Thái Nghe, Nguyễn Hùng Dũng: "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, trang 36-51. ISSN: 1859-2333, 2014.

[17]. Nguyen Thai-Nghe, Lars Schmidt-Thieme. 2015. Factorization Forecasting Approach for User Modeling. Journal of Computer Science and Cybernetics. 133-148. Vol 31, No 2. ISSN: 1813-9663. DOI: 10.15625/1813-9663/31/2/5860

[18]. Seung-Taek Park, Wei Chu. "Pairwise preference regression for cold-start recommendation". In SIGIR, pages 21-28. ACM, 2009.

[19]. Mingxuan Sun , Ke Zhou, Fuxin Li , Guy Lebanon, Joonseok Lee, Hongyuan Zha. "Learning Multiple-Question Decision Trees for Cold-Start Recommendation". In SIGIR, pages 445-454. ACM, 2013.

[20]. Mi Zhang, Jie Tang, Xuchen Zhang, Xiangyang Xue: "Addressing cold start in recommender systems: a semi-supervised co-training algorithm". In SIGIR, pages:73-82. ACM, 2014.

[21]. Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. "Functional matrix factorizations for coldstart recommendation". In SIGIR, pages 315–324. ACM, 2011

[22]. R. D. Snee: "Validation of Regression Models: Methods and Examples", Technometrics, Vol.19, No.4. (Nov,1977), pp. 415-428.

[23]. Budura, A., Michel, S., Cudré-Mauroux, P., Aberer, K.: Neighborhood-Based Tag Prediction. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 608–622. Springer, Heidelberg (2009).

[24]. Shakhnarovish, D., Indyk: Nearest-Neighbor Methods in Learning and Vision. The MIT Press (2005).

[25]. B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the Tenth International Conference on the World Wide Web (WWW 10), pp. 285-295, 2001.

AN APROACH FOR COLD-START PROBLEM IN RECOMMENDER SYSTEMS

Dinh The An Huy, Chau Le Sa Lin, Nguyen Huu Hoa, Nguyen Thai Nghe

ABSTRACT *— Recommender system (RS) is very successful in e-commerce. It recommends the list of items for users by using implicit or explicit feedback. Explicit feedback is based on reviews, ratings,... of the users in the past. Implicit feedback is based on the items that is chosen or viewed or clicked, etc. by the user. However, problem in RS is the new user (or new item) because the system has no feedback in the past, thus it can't create the list of items for that new user (or new item). That problem is called "cold-start problem" in the RS. In this paper, we propose an approach to mitigate the cold-start problem based on attributes of the new user (e.g., ages, gender, occupation...). Experiments are built to verify the feasibility of the proposed models.*

Keywords *— Recommender systems, cold-start problem, new user, new item.*