Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jsi.2017.004

TƯ VẤN LAI GHÉP DỰA TRÊN CÁC ĐỘ ĐO HÀM Ý THỐNG KÊ

Phan Phương Lan[1], Huỳnh Hữu Hưng[2] và Huỳnh Xuân Hiệp[1]
[1]Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ
[2]Khoa Công nghệ Thông tin, Trường Đại học Bách khoa, Đại học Đà Nẵng

Thông tin chung:
Ngày nhận bài: 15/09/2017
Ngày nhận bài sửa: 10/10/2017
Ngày duyệt đăng: 20/10/2017

Title:
Hybrid recommendation systems based on statistical implicative measures

Từ khóa:
Cường độ hàm ý, hệ tư vấn lai ghép, láng giềng gần, luật kết hợp, tính tiêu biểu

Keywords:
Implicative intensity, hybrid recommendation system, nearest neighbors, association rules, typicality

ABSTRACT

This paper proposes a hybrid recommendation model based on statistical implicative measures to suggest a list of top N items to an active user. The proposed model is built on two sub-models: the user-based collaborative filtering model and the association rule based model. The hybrid recommendation model is compared to its sub-models and some existing models such as latent factor model, popular model, and user-based collaborative filtering using Cosine on two datasets MSWeb and DKHP. The experimental results show that the performance of the proposed model is better than the compared models.

TÓM TẮT

Bài báo này đề xuất một mô hình tư vấn lai ghép dựa trên các độ đo hàm ý thống kê nhằm gợi ý cho người dùng danh sách các mục dữ liệu phù hợp. Mô hình đề xuất được xây dựng trên hai mô hình con: tư vấn lọc cộng tác dựa trên k láng giềng (người dùng) gần nhất và tư vấn dựa trên tập luật kết hợp. Mô hình tư vấn lai ghép được đánh giá trên hai tập dữ liệu MSWeb và DKHP khi so với các mô hình con của nó và một số mô hình tư vấn hiện có như: dựa trên nhân tố tiềm ẩn, dựa trên các mục dữ liệu phổ biến nhất, và lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine. Kết quả thực nghiệm cho thấy mô hình đề xuất có hiệu suất cao hơn so với các mô hình đó.

Trích dẫn: Phan Phương Lan, Huỳnh Hữu Hưng và Huỳnh Xuân Hiệp, 2017. Tư vấn lai ghép dựa trên các độ đo hàm ý thống kê. Tạp chí Khoa học Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin: 25-33.

1 GIỚI THIỆU

Hệ tư vấn (recommendation system/ recommender system - RS) (Ricci, 2011) là kỹ thuật hay công cụ phần mềm được nhúng trong một ứng dụng hoặc trang web để dự đoán sở thích của một cá nhân hoặc một nhóm người dùng đối với một sản phẩm hoặc dịch vụ cụ thể; và/hoặc giới thiệu các sản phẩm hoặc dịch vụ thích hợp cho một cá nhân hoặc một nhóm người dùng, qua đó giúp làm giảm tình trạng quá tải thông tin. Vì vậy, các hệ tư vấn được áp dụng trong nhiều lĩnh vực của cuộc sống (Lu *et al.*, 2015) như thương mại điện tử, dịch vụ điện tử... Các kỹ thuật (phương pháp) tư vấn được chia thành hai lớp chính (Aggarwal, 2016; Jannach *et al.*, 2011; Ricci, 2011): lớp các kỹ thuật cơ bản như lọc cộng tác, lọc nội dung hay dạng lai ghép; và lớp các kỹ thuật được phát triển trên nền những kỹ thuật cơ bản kết hợp với các thông tin bổ sung dựa trên ngữ cảnh hay mạng xã hội. Theo kỹ thuật tư vấn, các hệ tư vấn được phân thành nhiều nhóm khác nhau (Aggarwal, 2016; Jannach *et al.*, 2011; Ricci, 2011) dựa trên: nội dung, lọc cộng tác, nhân khẩu học, tri thức, lai

ghép, ngữ cảnh, mạng xã hội và theo nhóm. Trong các hướng nghiên cứu về hệ tư vấn, việc đề xuất các mô hình tư vấn mới hay cải tiến các phương pháp tư vấn hiện có là hướng nghiên cứu cốt lõi và nhận được nhiều quan tâm nhất.

Hệ tư vấn lai ghép (hybrid recommendation system) là hệ thống kết hợp hai hoặc nhiều kỹ thuật tư vấn để đạt được hiệu suất tốt hơn và để khắc phục những mặt hạn chế của từng kỹ thuật. Các hệ tư vấn lai có thể được phân thành bảy biến thể: trọng số, chuyển mạch, tầng, bổ sung đặc trưng, kết hợp đặc trưng, siêu mức (meta-level), và hỗn hợp (Aggarwal, 2016; Burke, 2007); và được thiết kế theo một trong ba dạng (Jannach *et al.*, 2011): nguyên khối, song song và tuần tự. Trong đó, thiết kế song song có ưu điểm là ít xâm lấn nhất đối với những thực thi hiện có. Cụ thể, các hệ tư vấn khác nhau hoạt động độc lập với nhau, và các dự đoán của từng hệ riêng lẻ được kết hợp vào lúc cuối.

Phân tích hàm ý thống kê (Statistical Implicative Analysis) (Gras *et al.*, 2009) là phương pháp phân tích dữ liệu dựa trên các độ đo hàm ý thống kê. Chúng được sử dụng để phát hiện các khuynh hướng giữa các thuộc tính dữ liệu (những mối quan hệ hàm ý mạnh), hay đo tính điển hình của một đối tượng đối với sự hình thành của một mối quan hệ, hay đo trách nhiệm của một đối tượng đối với sự tồn tại của một mối quan hệ. Vì vậy, những độ đo hàm ý thống kê có thể được sử dụng trong các hệ tư vấn.

Bài báo này đề xuất một mô hình tư vấn lai ghép mới dựa trên lọc cộng tác sử dụng k người dùng gần nhất và dựa trên tập luật kết hợp để gợi ý top N mục dữ liệu cho người cần được tư vấn. Dữ liệu đầu vào của mô hình là các đánh giá ở dạng nhị phân. Trong mô hình đề xuất, một số độ đo hàm ý thống kê được kết hợp thành những độ đo dùng để xếp hạng và lọc ra những mục phù hợp nhất.

Phần còn lại của bài báo được tổ chức thành bốn nội dung: trước tiên là sự mô tả ngắn gọn các độ đo cường độ hàm ý, tính trách nhiệm và tính tiêu biểu; tiếp theo là phần đề xuất mô hình tư vấn lai ghép dựa trên các độ đo đã đề cập ở trên; sau đó là phần trình bày và thảo luận các kết quả thực nghiệm; cuối cùng là phần kết luận.

2 CÁC ĐỘ ĐO HÀM Ý THỐNG KÊ

2.1 Độ đo cường độ hàm ý

Gọi E là một tập gồm n đối tượng được mô tả bởi một tập hữu hạn các thuộc tính nhị phân V. Gọi $A \subset E$ là một tập con gồm các đối tượng thỏa thuộc tính a (có thuộc tính a là đúng); Gọi $B \subset E$ là một tập con gồm các đối tượng thỏa thuộc tính b; $\bar{B}$ là phần bù của B; $n_a = card(A)$ là số phần tử của tập A; $n_b = card(B)$ là số phần tử của tập B, $n_{\bar{b}} = n - n_b = card(\bar{B})$ là số phần tử của tập $\bar{B}$; và $n_{a\bar{b}} = card(A \cap \bar{B})$ là số đối tượng thỏa thuộc tính a nhưng không thỏa thuộc tính b.

Gọi X và Y là hai tập ngẫu nhiên có số phần tử là n_a và n_b tương ứng. Biến ngẫu nhiên $card(X \cap \bar{Y})$ tuân theo phân phối Poisson với tham số $\lambda = \frac{n_a n_{\bar{b}}}{n}$. Xác suất của $card(X \cap \bar{Y}) \leq card(A \cap \bar{B})$ được xác định bởi (1).

$$Pr[card(X \cap \bar{Y}) \leq card(A \cap \bar{B})] = \sum_{s=0}^{card(A \cap \bar{B})} \frac{\lambda^s}{s!} e^{-\lambda} \quad (1)$$

Với $n_{\bar{b}} \neq 0$, biến ngẫu nhiên $card(X \cap \bar{Y})$ được chuyển về biến ngẫu nhiên được chuẩn hóa $Q(a, \bar{b})$ như (2). Trong thực nghiệm, giá trị quan sát của $Q(a, \bar{b})$ được biểu diễn bởi $q(a, \bar{b})$ và được định nghĩa theo (3). $q(a, \bar{b})$ được gọi là *chỉ số hàm ý* (Gras *et al.*, 2009).

$$Q(a, \bar{b}) = \frac{card(X \cap \bar{Y}) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} \quad (2)$$

$$q(a, \bar{b}) = \frac{n_{a\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} \quad (3)$$

Để đo khuynh hướng của mối quan hệ $a \rightarrow b$, độ đo *cường độ hàm ý* (Gras *et al.*, 2009) được xây dựng và có công thức tính như (4).

$$\varphi(a, b) = \begin{cases} 1 - \Pr\left(Q(a, \bar{b}) \leq q(a, \bar{b})\right) = \\ \frac{1}{2\pi} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt \text{ nếu } n_b \neq n \\ 0 \text{ ngược lại} \end{cases} \quad (4)$$

Mối quan hệ $a \rightarrow b$ là có thể chấp nhận với một ngưỡng α nếu $\varphi(a, b) \geq 1 - \alpha$.

2.2 Độ đo trách nhiệm

Để đo sự đóng góp của một đối tượng i đối với sự tồn tại của mối quan hệ $a \rightarrow b$, *độ đo trách nhiệm* $\varphi_{i,a \rightarrow b}$(Gras *et al.*, 2009) được phát triển và xác định theo (5). Trong đó, $a(i)$ (resp. $b(i)$) là giá trị nhị phân cho biết có sự hiện diện hay không có sự hiện diện của thuộc tính a (resp. b) trong đối tượng i.

$$\varphi_{i,a\to b} = 1 \text{ nếu } [a(i) = 1 \text{ hay } a(i) = 0] \text{ và } b(i) = 1$$
$$\varphi_{i,a\to b} = 0 \text{ nếu } a(i) = 1 \text{ và } b(i) = 0 \quad (5)$$
$$\varphi_{i,a\to b} = p \in]0,1[\text{ nếu } a(i) = b(i) = 0$$

Trong thực tế, p được đặt bằng 0.5.

2.3 Độ đo tính tiêu biểu

Để đo tính tiêu biểu của một đối tượng i trong sự hình thành mối quan hệ $a \to b$, *độ đo tính tiêu biểu* (Gras *et al.*, 2009) được đề xuất và tính như công thức (6).

$$\gamma(i, a \to b) = 1 - \frac{d_1(i, a \to b)}{max_{j\in E} d_1(j, a \to b)} \quad (6)$$

Trong đó, $d_1(i, a \to b)$ là khoảng cách giữa đối tượng i và mối quan hệ $a \to b$. Giá trị của $d_1(i, a \to b)$ được xác định theo (7) với $\varphi(a, b)$ là cường độ hàm ý của $a \to b$ và $\varphi_{i,a\to b}$ là trách nhiệm của đối tượng i đối với sự hình thành mối quan hệ $a \to b$.

$$d_1(i, a \to b) = \left(\frac{(\varphi(a,b) - \varphi_{i,a\to b})^2}{1 - \varphi(a,b)}\right)^{1/2} \quad (7)$$

3 MÔ HÌNH TƯ VẤN LAI GHÉP DỰA TRÊN CÁC ĐỘ ĐO HÀM Ý THỐNG KÊ

Mô hình tư vấn (hay mô hình hệ tư vấn) lai ghép đề xuất có thể được hình thức hóa như sau:

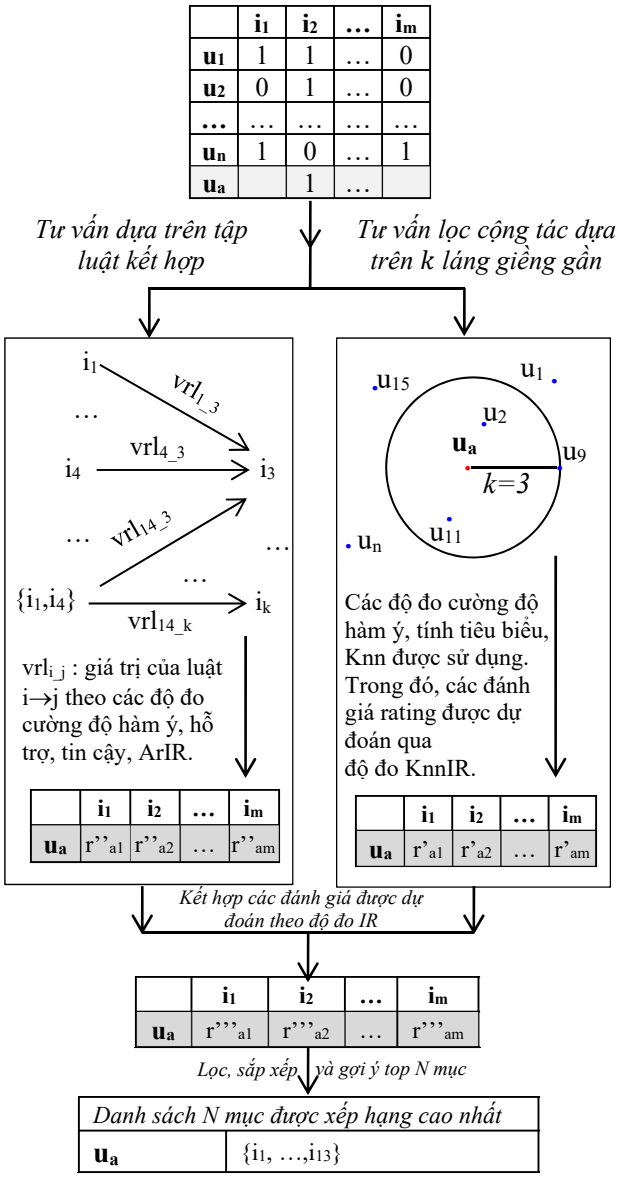
– Gọi $U = \{u_1, u_2, \dots, u_n\}$ là một tập hữu hạn những người dùng.

– Gọi $I = \{i_1, i_2, \dots, i_m\}$ là tập hữu hạn các mục dữ liệu (mục, item). Mục có thể là bài hát, bộ phim, mặt hàng...

– Gọi $R = (r_{jk})$ với $j = 1..n$ và $k = 1..m$ là một ma trận đánh giá (ma trận xếp hạng, rating matrix) lưu trữ thông tin phản hồi của người dùng về các mục dữ liệu. $r_{jk} = 0$ nếu mục dữ liệu i_k không được thích (không được cần hay không được biết) bởi người dùng u_j; $r_{jk} = 1$ nếu mục dữ liệu i_k được thích bởi người dùng u_j.

– Gọi $f: U \times I \to R$ là một hàm ánh xạ từ những kết hợp của người dùng và các mục vào các đánh giá r.

Mục tiêu của mô hình đề xuất là tìm một hàm $f': U \times I \to R'$ sao cho hàm $\xi(r, r')$ đạt hiệu suất cao hơn (qua các độ đo precision, recall và accuracy) trong việc gợi ý top N mục dữ liệu phù hợp cho người dùng.



Hình 1: Bản phác thảo mô hình tư vấn lai ghép dựa trên các độ đo hàm ý thống kê

Hình 1 là bản phác thảo mô hình tư vấn lai ghép được đề xuất. Trong đó, mô hình tư vấn con dựa trên tập luật kết hợp sử dụng độ đo $ArIR$ được kết hợp từ cường độ hàm ý ($\varphi(rule)$) và độ tin cậy ($Conf$) để dự đoán các đánh giá; mô hình tư vấn con dựa trên k người dùng gần nhất sử dụng độ đo $KnnIR$ được xây dựng từ độ đo tính tiêu biểu γ để dự đoán các đánh giá. Mô hình tư vấn lai ghép sau đó sẽ kết hợp các đánh giá được dự đoán từ hai mô hình con theo độ đo IR. Công thức tính của $ArIR$, $KnnIR$, IR được lần lượt trình bày trong (8), (10) và (12).

$$ArIR(u_a, i) = \frac{AIR(u_a, i)}{max_{l \in I} AIR(u_a, l)} \quad (8)$$

với

$$AIR(u_a, i) = \sum_{j=1}^{r} Conf(rule_j) * \varphi(rule_j) \quad (9)$$

nếu u_a đã đánh giá các mục ở bên vế trái của luật $rule_j$ và i nằm bên vế phải của luật $rule_j$. r là số luật của tập luật.

$$KnnIR(u_a, i) = \frac{KIR(u_a, i)}{max_{l \in I} KIR(u_a, l)} \quad (10)$$

với

$$KIR(u_a, i) = \sum_{j=1}^{k} \gamma(i, u_a \to u_j) \quad (11)$$

nếu u_j đã đánh giá các mục i. k là số láng giềng gần nhất của u_a.

$$IR(u_a, i) = \frac{1}{2} * (ArIR(u_a, i) + KnnIR(u_a, i)) \quad (12)$$

Mô hình lai ghép sẽ sử dụng Giải thuật 1 để gợi ý top N mục dữ liệu cho người dùng. Trong quá trình tư vấn, mô hình sẽ sử dụng thêm các Giải thuật 2, Giải thuật 3 và Giải thuật 4.

3.1 Tư vấn lai ghép dựa trên các độ đo hàm ý thống kê

Giải thuật 1 (tư vấn lai ghép dựa trên các độ đo hàm ý thống kê *HybridImplicativeRS*) có:

– Input gồm ma trận đánh giá R với tập người dùng U và tập mục dữ liệu I; vector A có m phần tử với $given$ đánh giá đã biết của người cần được tư vấn u_a; số láng giềng gần nhất k; các ngưỡng của độ hỗ trợ s, độ tin cậy c và cường độ hàm ý ii.

– Output là một danh sách gồm N mục dữ liệu có xếp hạng cao nhất được gợi ý cho u_a.

Các bước thực hiện gồm: (1) - Dự đoán đánh giá/xếp hạng của người dùng u_a cho từng mục dữ liệu $i_j \in I$ dựa trên k láng giềng gần nhất; (2) - Dự đoán đánh giá của người dùng u_a cho từng mục dữ liệu $i_j \in I$ dựa trên tập luật kết hợp; (3) - Tổng hợp giá trị đánh giá của người dùng u_a cho từng mục dữ liệu $i_j \in I$ theo (12); (4) - Loại bỏ $given$ mục dữ liệu đã biết trước của u_a ra khỏi danh sách các mục được dự đoán đánh giá; (5) - Sắp xếp và gợi ý N mục dữ liệu được xếp hạng cao nhất cho u_a.

```
HybridImplicativeRS(vector A; ratingmatrix R; int k; float s,c,ii){
    KnnIR = KnnImplicativeRS(A,R,k);
    ArIR = ArImplicativeRS(A,R,s,c,ii);
    IR = f(KnnIR,ArIR);
    Ratings = RemoveKnownGiven(A, IR);
    TopNItems(Ratings);    }
```

3.2 Tư vấn lọc cộng tác dựa trên *k* người dùng gần nhất

Giải thuật 2 (tư vấn lọc cộng tác dựa trên k người dùng gần nhất *KnnImplicativeRS*) có:

– Input gồm ma trận đánh giá R với tập người dùng U và tập mục dữ liệu I; vector A có m phần tử với $given$ đánh giá đã biết của người cần được tư vấn u_a; số láng giềng gần nhất k.

– Output gồm các đánh giá của người dùng u_a cho các mục dữ liệu $i_j \in I$.

– Các bước thực hiện gồm: (1) - Đo cường độ hàm ý của mối quan hệ giữa hai người dùng (u_a, u_i) với $u_i \in U$ theo giải thuật đã được chúng tôi trình bày trong (Phan *et al.*, 2016b); (2) - Tìm k láng giềng gần nhất của u_a có cường độ hàm ý cao nhất; (3) - Tính giá trị tiêu biểu *Typicality* của mỗi mục dữ liệu $i_j \in I$ đối với sự hình thành mối quan hệ (u_a, u_l) với u_l là một trong k láng giềng gần nhất của u_a theo Giải thuật 3; (4) - Dự đoán đánh giá của người dùng u_a cho từng mục dữ liệu $i_j \in I$.

```
KnnImplicativeRS (vector A; ratingmatrix R; int k){
  IIntensity = CalculateImplicativeIntensity(A,R);
  Neighbors = KNearestNeighbors(IIntensity, k);
  Typic =  Typicality(A,Neighbors,R,IIntensity);
  for each i_j ∈ I do
    KIR[u_a, i_j] = colSum(Typic * R[Neighbors, i_j]);
  KnnIR[u_a, ] = KIR[u_a, ]/max(KIR[u_a, ];
  return KnnIR;        }
```

3.3 Đo tính tiêu biểu của một mục dữ liệu đối với sự hình thành của một quan hệ

Giải thuật 3 (đo tính tiêu biểu của một mục dữ liệu đối với sự hình thành của một quan hệ - *Typicality*) có:

– Input gồm ma trận đánh giá R với tập người dùng U và tập mục dữ liệu I; vector A có m phần tử với $given$ đánh giá đã biết của người cần được tư vấn u_a; vector $Neighbors$ chứa k láng giềng gần nhất của u_a; và vector $IIntensity$ chứa cường độ hàm ý của mối quan hệ (u_a, u_l) với $u_l \in Neighbors$.

– Output gồm các giá trị tiêu biểu của mục dữ liệu $i_j \in I$ đối với sự hình thành mối quan hệ $(u_a, u_l), u_l \in Neighbors$.

– Các bước thực hiện gồm: (1) - Đo tính trách nhiệm của mục dữ liệu $i_j \in I$ đối với sự tồn tại của mối quan hệ (u_a, u_l) với $u_l \in Neighbors$ theo công thức (5); (2) - Tính khoảng cách từ mục i_j đến mối quan hệ (u_a, u_l) theo (7); (3) - Tìm giá trị tiêu biểu của mục i_j đối với sự hình thành mối quan hệ (u_a, u_l) theo (6).

Typicality(vector A,vector Neighbors, ratingmatrix R, vector IIntensity){
  for each $u_l \in$ Neighbors do
    for each item $i_j \in$ I do
      if (R[u_l, i_j]=1) Contribution[u_l, i_j]=1;
      else if (A[i_j]=1 and R[u_l, i_j]=0)
        Contribution[u_l, i_j]=0;
      else if (A[i_j]=0 and R[u_l, i_j]=0)
        Contribution[u_l, i_j]=0.5;
  ua_intensity = IIntensity [Neighbors];
  for each item $i_j \in$ I do {
    ua_contribution = Contribution[Neighbors, i_j];
    Dist[Neighbors, i_j] = sqrt((ua_intensity-
        ua_contribution)^2/ (1-ua_intensity));
  }
  Typicality = Null;
  for each $u_i \in$ Neighbors do {
    Rowmax = max(Dist[u_i,]);
    Typic = 1 – Dist[u_i,] / Rowmax;
    Typicality = rowbind(Typicality,Typic); }
    return Typicality;}

3.4 Tư vấn lọc cộng tác dựa trên tập luật kết hợp

Giải thuật 4 (tư vấn lọc cộng tác dựa trên tập luật kết hợp *ArImplicativeRS*) có:

– Input gồm ma trận đánh giá R với tập người dùng U và tập mục dữ liệu I; vector A có m phần tử với $given$ đánh giá đã biết của người cần được tư vấn u_a; các ngưỡng của độ hỗ trợ s, độ tin cậy c và cường độ hàm ý ii.

– Output là các đánh giá của người dùng u_a cho các mục dữ liệu $i_j \in I$.

– Các bước thực hiện gồm: (1) - Sinh tập luật kết hợp bằng giải thuật Apriorio, sử dụng các ngưỡng hỗ trợ, tin cậy để cắt tập luật; (2) - Tính cường độ hàm ý cho từng luật trong tập luật theo giải thuật được chúng tôi trình bày trong (Phan *et al.*, 2016a) và lọc lại tập luật dựa vào ngưỡng cường độ hàm ý; (3) - Dự đoán đánh giá của người dùng u_a cho từng mục dữ liệu dựa trên tập luật kết hợp đã được lọc.

ArImplicativeRS (vector A; ratingmatrix R; float s,c,ii) {
  Ruleset = Apriorio(R,s,c);
  Ruleset = ImplicativeIntensity(Ruleset,R,ii);
  Ruleuserleft = Subset(LeftHandSide(Ruleset),A);
  Rules = NULL;
  for each item $i_j \in$ I
    for each rule r ∈ Ruleset
      if (Ruleuserleft[r, u_a] = 1) and
        (Righthandside[i_j,r]=1)
        AIR[u_a, i_j]= AIR[u_a, i_j]+
        Ruleset[r,Conf]* Ruleset[r,IIntensity];
  ArIR[u_a,] = AIR[u_a,]/max(AIR[u_a,]
  return ArIR; }

4 THỰC NGHIỆM

4.1 Thiết lập thực nghiệm

4.1.1 Dữ liệu thực nghiệm

Hai tập dữ liệu được sử dụng là MSWeb (Asuncion and Newman, 2007) và DKHP. Tập dữ liệu MSWeb được tạo ra bằng cách lấy mẫu và xử lý các nhật ký (log) của www.microsoft. com trong khoảng thời gian một tuần. Với từng người dùng trong số những người dùng ẩn danh và được chọn ngẫu nhiên, tập dữ liệu này lưu tất cả các mục của trang web (Vroots) được truy cập bởi người đó. MSWeb gồm: 32710 người dùng, 285 Vroot và 98653 đánh giá với giá trị TRUE. Tập dữ liệu DKHP được thu thập thông qua website đăng ký học phần của trường Đại học Cần Thơ (https://htql.ctu.edu.vn). Tập dữ liệu này lưu kết quả đăng ký học phần cho học kỳ thứ ba (trong chín học kỳ[1]) của sinh viên khóa 40 và 41 thuộc Khoa Công nghệ thông tin và Truyền thông. Tập dữ liệu chứa: 1.172 sinh viên, 81 học phần và 5.705 đánh giá (đăng ký) với giá trị TRUE.

Bảng 1: Thông tin chung của hai tập dữ liệu MSWeb và DKHP sau khi được lọc

Tập dữ liệu	Số người dùng	Số mục dữ liệu	Số đánh giá	Số given tối đa[2]
MSWeb	875	135	10.487	7
DKHP	779	36	4.095	3

Để tăng tính chính xác trong việc đưa ra gợi ý, các tập dữ liệu thực nghiệm cần được tiền xử lý. Nếu ta giữ các mục chỉ được đánh giá vài lần và những người chỉ đánh giá một vài mục thì các đánh giá có thể bị thiên vị. Bên cạnh đó, khi thực hiện tư vấn trên tập dữ liệu DKHP, ta cũng cần lưu ý đến

[1] Chương trình học trong 4.5 năm (9 học kỳ).
[2] Số lượng mục dữ liệu tối đa được chọn ngẫu nhiên trên mỗi người dùng trong tập kiểm thử được sử dụng để xây dựng các gợi ý và đánh giá mô hình đề xuất; số lượng này được xác định dựa trên phân vị.

những ràng buộc trong quy định đăng ký học phần[3]. Kết quả, với tập dữ liệu MSWeb, số người dùng đã xem ít nhất 10 mục website (số Vroot) và số mục website được xem bởi ít nhất 50 người được chọn để trích xuất dữ liệu. Với tập dữ liệu DKHP, số sinh viên đã đăng ký từ 5 học phần trở lên và số học phần được đăng ký bởi ít nhất 25 sinh viên được chọn để trích xuất dữ liệu. Thông tin chung về hai tập dữ liệu MSWeb và DKHP sau khi lọc được trình bày trong Bảng 1.

4.1.2 Công cụ thực nghiệm

Mô hình tư vấn đề xuất được cài đặt bằng ngôn ngữ R và các hàm của công cụ Interestingnesslab (đã được chúng tôi phát triển trong (Phan *et al.*, 2017a)). Bên cạnh đó, chúng tôi sử dụng một số mô hình tư vấn của gói recommenderlab (Hahsler, 2011) để so sánh với mô hình đề xuất như: mô hình tư vấn dựa trên nhân tố tiềm ẩn, mô hình tư vấn dựa trên các mục dữ liệu phổ biến nhất, và mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine.

4.1.3 Đánh giá mô hình tư vấn

Để đánh giá mô hình tư vấn, tập dữ liệu đầu vào được phân tách thành tập dữ liệu huấn luyện và tập dữ liệu kiểm thử. Tập dữ liệu kiểm thử lại được chia thành tập dữ liệu truy vấn và tập dữ liệu đích có cùng kích thước. Trong đó, với mỗi người dùng, tập dữ liệu truy vấn chỉ có $given$ đánh giá được chọn ngẫu nhiên; tập dữ liệu đích gồm những đánh giá còn lại. Tập dữ liệu truy vấn cùng với tập dữ liệu huấn luyện được sử dụng để dự đoán các giá trị xếp hạng trong khi tập dữ liệu đích được sử dụng để đánh giá kết quả gợi ý. Phương pháp k-fold cross validation được áp dụng để phân tách tập dữ liệu thành k tập con có kích thước bằng nhau và thực hiện k lần đánh giá sau đó lấy kết quả trung bình. Ở mỗi lần đánh giá, $(k-1)$ tập con được sử dụng làm tập huấn luyện và 1 tập con còn lại được sử dụng làm tập kiểm thử. Trong thực nghiệm này, k được chọn là 4.

Do mô hình đề xuất sử dụng dữ liệu đầu vào ở dạng nhị phân và xuất ra (gợi ý) cho người dùng danh sách các mục dữ liệu phù hợp nên các độ đo dùng để đánh giá sự dự đoán rating như RMSE, MAE không thực sự phù hợp (Gunawardana and Shani, 2009).

Khi số lượng các mục dữ liệu (độ dài của danh sách gợi ý) cần được giới thiệu đến người dùng không được xác định trước, việc đánh giá thuật toán trên một dải các độ dài của danh sách gợi ý sẽ thích hợp hơn là sử dụng một độ dài cố định. Vì vậy, các đường cong Precision - Recall và ROC (Receiver Operating Characteristic) thường được sử dụng (Gunawardana and Shani, 2009). Ngoài ra, đường cong ROC thường được sử dụng để so sánh hiệu suất của nhiều giải thuật tư vấn. Đường cong ROC của giải thuật nào nằm trên hoàn toàn các đường cong ROC của những giải thuật khác thì hiệu suất của giải thuật đó là tốt hơn. Đường cong Precision - Recall được xây dựng theo: độ chính xác (precision) và độ bao phủ (recall). Đường cong ROC được xây dựng theo: độ nhạy và phần bù của độ đặc hiệu. Độ nhạy (còn gọi là True Positive Rate - TPR) được tính như độ bao phủ. Phần bù của độ đặc hiệu còn có tên là False Positive Rate (FPR). Những phép đo này được xây dựng dựa vào ma trận nhầm lẫn như Bảng 2 và có công thức tính như (13), (14) và (15). Ngoài ra, độ đo chuẩn xác accuracy cũng được sử dụng khi đánh giá các mô hình tư vấn và được tính theo công thức (16).

Bảng 2: Ma trận nhầm lẫn

Thực tế/Gợi ý	Không được gợi ý	Được gợi ý
Không được thích	True–Nagative TN	False–Positive FP
Được thích	False–Nagative FN	True–Positive TP

$$precision = \frac{TP}{TP + FP} \quad (13)$$

$$recall/TPR = \frac{TP}{TP + FN} \quad (14)$$

$$FPR = \frac{FP}{TN + FP} \quad (15)$$

$$accuracy = \frac{TN + TP}{TN + FP + FN + TP} \quad (16)$$

4.2 Kết quả thực nghiệm

4.2.1 Đánh giá mô hình tư vấn đề xuất với các mô hình con của nó

Trong mô hình tư vấn dựa trên tập luật, để không bỏ sót các luật kết hợp với chất lượng cao, ngưỡng hỗ trợ và ngưỡng tin cậy nên được gán các giá trị nhỏ. Tuy nhiên, ta cần chọn các ngưỡng phù hợp để giảm kích thước của tập luật (qua đó làm giảm thời gian tư vấn) mà vẫn đảm bảo hiệu suất (như tính chính xác và bao phủ) khi đưa ra gợi ý. Việc chọn các ngưỡng này đã được trình bày trong (Phan *et al*., 2017b).

Bảng 3 trình bày kết quả gợi ý trung bình của 50 lần thực thi bằng phương pháp k-fold cross validation trên tập dữ liệu MSWeb; số đánh giá (số Vroot) biết trước của mỗi người dùng trong tập kiểm thử $given = 7$; số láng giềng gần nhất của mỗi người dùng trong tập kiểm thử $k = 30$; các

[3] Sinh viên được đăng ký học tối đa 20 tín chỉ trong một học kỳ, trừ học kỳ cuối cùng (tối đa: 25 tín chỉ). Trung bình, mỗi học phần khoảng 3 tín chỉ. Điều kiện để mở lớp học phần: số sinh viên đăng ký >= 25. Điều kiện để xét cấp học bổng: sinh viên phải đăng ký >= 15 tín chỉ.

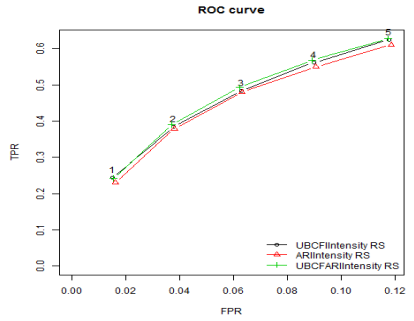
ngưỡng của độ đo hỗ trợ, tin cậy và cường độ hàm ý lần lượt là: $s = 0.01$, $c = 0.1$ và $ii = 0.5$. Mô hình lai ghép và hai mô hình con của nó (dựa trên những người dùng gần nhất và dựa trên tập luật) cần gợi ý 1, 3, 6, 9, 12, 15 Vroot cho người dùng.

Bảng 3: Kết quả so sánh ba mô hình tư vấn trên tập dữ liệu MSWeb

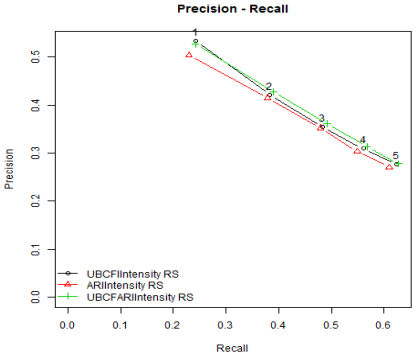
Số mục gợi ý	Precision	Recall/ TPR	FPR	Accuracy
Mô hình tư vấn dựa trên 30 người dùng gần nhất (UBCFIIntensity RS)				
1	0,563824	0,130065	0,003523	0,961970
3	0,462436	0,309841	0,013038	0,959212
6	0,364491	0,474111	0,030860	0,948269
9	0,296350	0,567219	0,051297	0,932335
12	0,248971	0,629189	0,073049	0,913905
15	0,214884	0,674067	0,095501	0,894149
Mô hình tư vấn dựa tập luật kết hợp (ARIIntensity RS)				
1	0,560837	0,129018	0,003548	0,961923
3	0,457700	0,304850	0,013151	0,958990
6	0,355121	0,458077	0,031313	0,947390
9	0,290573	0,552236	0,051714	0,931522
12	0,245807	0,616188	0,073349	0,913312
15	0,212706	0,661517	0,095757	0,893638
Mô hình tư vấn lai ghép (UBCFARIIntensity RS)				
1	0,575385	0,132997	0,003429	0,962151
3	0,470965	0,315472	0,012828	0,959612
6	0,365735	0,473481	0,030794	0,948386
9	0,297878	0,567486	0,051178	0,932549
12	0,250997	0,630861	0,072842	0,914285
15	0,216968	0,676788	0,095236	0,894637

Kết quả thực nghiệm cho thấy mô hình lai ghép có giá trị chính xác precision, bao phủ recall cao hơn cũng như có tỷ lệ số Vroot được hệ thống gợi ý nhưng không được người dùng xem trên tổng số Vroot được gợi ý (FPR) là thấp hơn so với hai mô hình con của nó. Ngoài ra, tỷ lệ số gợi ý chính xác trên tổng số gợi ý có thể có (giá trị accuracy) của mô hình đề xuất cũng cao hơn hai mô hình con. Khi thay đổi các thông số: số lần thực thi, số $given$ và số láng giềng k, ta cũng nhận được kết quả: hiệu suất của mô hình lai ghép cao hơn so với hai mô hình con của nó.

Hình 2 và Hình 3 hiển thị đường cong ROC và đường cong Precision - Recall của ba mô hình tư vấn trên tập dữ liệu DKHP trong 30 lần thực thi với số học phần cần gợi ý cho sinh viên lần lượt là 1, 2, 3, 4, 5. Các thông số được sử dụng trong những mô hình này là: $given = 3$ và $k = 30$. Hình 2 và Hình 3 cho thấy khi số học phần cần gợi ý cho sinh viên là 1, mô hình lai ghép có giá trị chính xác và bao phủ xấp xỉ các giá trị của mô hình lọc cộng tác dựa trên người dùng và cao hơn của mô hình dựa trên tập luật. Khi cần gợi ý từ 2 học phần trở lên, hiệu suất của mô hình lai ghép tốt hơn hai mô hình con của nó. Khi thay đổi các thông số $given$ và k, ta cũng nhận được kết quả tương tự như Hình 2 và Hình 3.



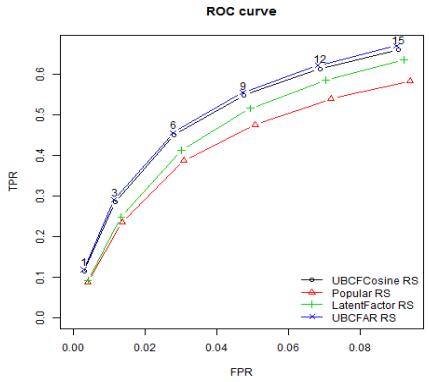
Hình 2: Đường cong ROC của ba mô hình tư vấn trên tập dữ liệu DKHP



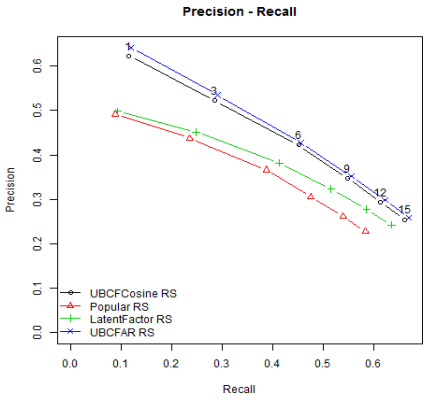
Hình 3: Đường cong Precision – Recall của ba mô hình tư vấn trên tập dữ liệu DKHP

4.2.2 Đánh giá mô hình đề xuất với một số mô hình tư vấn hiện có

Mô hình tư vấn lai ghép dựa trên các độ đo hàm ý thống kê (UBCFAR RS) được so sánh với ba mô hình tư vấn tích hợp sẵn trong gói recommenderlab: dựa trên nhân tố tiềm ẩn (LatentFactor RS), dựa trên các mục dữ liệu phổ biến nhất (Popular RS), và lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine (UBCFCosine RS).



Hình 4: Đường cong ROC của bốn mô hình tư vấn trên tập dữ liệu MSWeb



Hình 5: Đường cong ROC của bốn mô hình tư vấn trên tập dữ liệu MSWeb

Hình 4 và Hình 5 hiển thị đường cong ROC và đường cong Precision - Recall của bốn mô hình tư vấn trên tập dữ liệu MSWeb trong 20 lần thực thi, $given = 6$, $k = 30$, $s = 0.01$, $c = 0.1$ và $ii = 0.5$ với số Vroot cần gợi ý cho người dùng lần lượt là 1, 3, 6, 9, 12 và 15. Khi thay đổi các thông số: số lần thực thi, $given$ và k, ta cũng nhận được các kết quả tương tự như hai hình trên. Hình 4 cho thấy đường cong ROC của mô hình đề xuất hoàn toàn nằm trên các đường cong ROC còn lại. Bên cạnh đó, Hình 5 cho thấy giá trị chính xác và bao phủ của mô hình đề xuất cũng cao hơn so với ba mô hình còn lại trên từng số lượng Vroot cần gợi ý cho người dùng. Như vậy, hiệu suất của mô hình tư vấn lai ghép là cao hơn.

Bảng 4 trình bày kết quả gợi ý trung bình của 20 lần thực thi trên tập dữ liệu DKHP với $given = 2$, $k = 30$. Kết quả thực nghiệm cho thấy khi cần gợi ý 1 học phần cho sinh viên, mô hình tư vấn lai ghép cho giá trị chính xác và chuẩn xác thấp hơn một chút so với mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine và mô hình tư vấn dựa trên các mục phổ biến nhất. Tuy nhiên, khi cần gợi ý từ 2 học phần trở lên, mô hình lai ghép cho kết quả tốt hơn. Khi thay đổi các thông số: số lần thực thi, $given$ và k, ta cũng nhận được các kết quả tương tự. Ngoài ra, do quy định về số tín chỉ được đăng ký, hệ tư vấn sẽ hữu ích hơn nếu gợi ý cho sinh viên nhiều hơn 1 học phần. Vì vậy, hiệu suất của mô hình lai ghép là tốt hơn so với ba mô hình còn lại.

Bảng 4: Kết quả so sánh bốn mô hình tư vấn trên tập dữ liệu DKHP

Số mục gợi ý	Precision	Recall/ TPR	FPR	Accuracy
Mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine (UBCFCosine RS)				
1	0,645114	0,200794	0,011496	0,912836
2	0,522430	0,325782	0,030997	0,906939
3	0,453025	0,423452	0,053283	0,896010
4	0,400746	0,498849	0,077858	0,880946
5	0,361244	0,561906	0,103760	0,863489
Mô hình tư vấn dựa trên các mục phổ biến nhất (Popular RS)				
1	0,649619	0,201156	0,011339	0,913101
2	0,50514	0,310889	0,032094	0,904904
3	0,43805	0,404541	0,054711	0,893367
4	0,386374	0,476901	0,079706	0,877564
5	0,34448	0,533092	0,106480	0,858559
Mô hình tư vấn dựa trên nhân tố tiềm ẩn (LatentFactor RS)				
1	0,373033	0,118329	0,020384	0,896831
2	0,339689	0,214064	0,042930	0,885440
3	0,311337	0,293334	0,067159	0,871006
4	0,286818	0,359365	0,092734	0,854139
5	0,266701	0,417058	0,119189	0,835682
Mô hình tư vấn lai ghép (UBCFAR RS)				
1	0,625761	0,195686	0,012133	0,911698
2	0,525412	0,327471	0,030802	0,907289
3	0,460258	0,429969	0,052569	0,897287
4	0,406551	0,506183	0,077098	0,882312
5	0,365089	0,567467	0,103127	0,864620

5 KẾT LUẬN

Bài báo đã đề xuất một mô hình tư vấn lai ghép được phát triển trên hai mô hình tư vấn con (lọc cộng tác dựa trên người dùng, tập luật kết hợp) và một số độ đo hàm ý thống kê quan trọng nhằm gợi ý cho người dùng các mục dữ liệu phù hợp. Mô hình đề xuất sử dụng: (1) - dữ liệu đầu vào ở dạng nhị phân; (2) - độ đo cường độ hàm ý để tìm những láng giềng gần nhất của người cần được tư vấn và lọc tập luật; (3) - độ đo $ArIR$ được kết hợp từ cường độ hàm ý và độ tin cậy, độ đo $KnnIR$ được

xây dựng dựa trên tính tiêu biểu, và độ đo IR được hình thành từ $ArIR$ và $KnnIR$ để dự đoán các giá trị đánh giá/xếp hạng của người cần được tư vấn cho các mục dữ liệu trong từng mô hình con và mô hình lai ghép. Các kết quả thực nghiệm trên tập dữ liệu chuẩn MSWeb và tập dữ liệu thực DKHP cho thấy mô hình lai ghép đề xuất có hiệu suất tốt hơn so với các mô hình con của nó cũng như so với các mô hình tư vấn được tích hợp trong gói recommenderlab như: nhân tố tiềm ẩn, dựa trên các mục dữ liệu phổ biến nhất, và lọc cộng tác dựa trên người dùng sử dụng độ đo Cosine.

TÀI LIỆU THAM KHẢO

Aggarwal C., 2016. Recommender Systems: The Textbook. Springer International Publishing Switzerland, 498 pages.

Asuncion A. and Newman D.J., 2007. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science. http://www.ics.uci.edu/ ~mlearn/MLRepository.html.

Gras R., Suzuki E., Guillet F. and Spagnolo F., 2008. Statistical Implicative Analysis, Springer-Verlag, 513 pages.

Gunawardana A. and Shani G., 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. Journal of Machine Learning Research. 10: pp. 2935–2962.

Hahsler M., 2011. recommenderlab: A Framework for Developing and Testing Recommendation Algorithms, https://cran.r-project.org/web/packages/recommenderlab/index.html.

Jannach D., Zanker M., Felfernig A., and Friedrich G., 2011. An introduction to recommender systems. Cambridge University Press, 335 pages.

Lu J., Wu D., Mao M., Wang W., and Zhang G., 2015. Recommender system application developments: a survey. Decision Support Systems. 74: pp. 12-32.

Phan L.P., Nguyen K.M., Huynh H.H., and Huynh H.X., 2016a. Association-based recommender system using statistical implicative cohesion measure. In: Eighth International Conference on Knowledge and Systems Engineering, Hanoi, pp.144-149.

Phan Phương Lan, Trần Uyên Trang, Huỳnh Hữu Hưng, Huỳnh Xuân Hiệp, 2016b. Tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo gắn kết hàm ý thống kê. In: Hội nghị khoa học công nghệ quốc gia lần thứ IX. DOI: 10.1562/vap.2016.00093.

Phan L.P., Phan N.Q., Nguyen K.M., Huynh H.H., Huynh H.X., and Guillet F., 2017a. Interestingnesslab: A Framework for Developing and Using Objective Interestingness Measures. In: Advances in Intelligent Systems and Computing, Springer, 538: pp.302-311.

Phan Phương Lan, Huỳnh Hữu Hưng, Huỳnh Xuân Hiệp, 2017b. Hệ tư vấn dựa trên cường độ hàm ý và trách nhiệm. In: Hội nghị khoa học công nghệ quốc gia lần thứ X (accepted).

R.Burke, 2007. Hybrid Web Recommender Systems. The Adaptive Web: Methods and Strategies of Web Personalization, Springer -Verlag Berlin, Heidelberg, pp.377-408.

Ricci F., Rokach L., Shapira B., and Kantor P.B., 2011. Recommender Systems Handbook. Springer US, 842 pages.