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.2016.00029

HỆ TƯ VẤN LỌC CỘNG TÁC THEO NGƯỜI DÙNG DỰA TRÊN ĐỘ ĐO HÀM Ý THỐNG KÊ

Phan Quốc Nghĩa[1], Nguyễn Minh Kỳ[2], Đặng Hoài Phương[3], Huỳnh Xuân Hiệp[4,5]

[1] Phòng Khảo thí, Trường Đại học Trà Vinh
[2] Khoa Công nghệ Thông tin, Trường Đại học Kỹ thuật – Công nghệ Cần Thơ
[3] Khoa Công nghệ Thông tin, Trường Đại học Bách khoa, Đại học Đà Nẵng
[4] Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ
[5] Nhóm nghiên cứu liên ngành DREAM-CTU/IRD, Trường Đại học Cần Thơ

nghiatvnt@tvu.edu.vn, nmky@ctuet.edu.vn, dhphuong@dut.udn.vn, hxhiep@ctu.edu.vn

TÓM TẮT— *Từ khi ra đời đến nay, hệ tư vấn lọc cộng tác nói chung, hệ tư vấn lọc cộng tác dựa trên người dùng nói riêng đã có một sự phát triển vượt bậc về mặt ứng dụng kỹ thuật, công nghệ cũng như ứng dụng vào thực tế cuộc sống. Đặc biệt, hệ tư vấn được các nhà quản lý sử dụng làm công cụ hỗ trợ hữu hiệu trong nhiều lĩnh vực kinh doanh như Amazon, Netflix và Pandora. Tuy nhiên, các thế hệ hiện tại của hệ tư vấn vẫn chưa đáp ứng đầy đủ các yêu cầu của người sử dụng. Trong bài viết này chúng tôi đề xuất một tiếp cận mới cho hệ tư vấn lọc cộng tác dựa trên người dùng. Hệ tư vấn lọc cộng tác theo người dùng được xây dựng dựa trên độ đo hàm ý thống kê. Trong hệ tư vấn này, chúng tôi xây dựng một độ đo tương đồng dựa trên độ đo chỉ số hàm ý thống kê gọi là độ đo tương đồng hàm ý thống kê để xác định sự tương đồng giữa hai người dùng trong hệ thống. Thông qua thực nghiệm trên hai tập dữ liệu MovieLense và MSWeb cho thấy rằng độ đo tương đồng mà chúng tôi đề xuất cho kết quả khá tốt trên hệ tư vấn lọc cộng tác dựa trên người dùng so với các độ đo tương đồng truyền thống như Pearson correlation, Cosine similarity và Jaccard.*

Từ khóa— *Độ đo tương đồng, độ đo chỉ số hàm ý thống kê, hệ tư vấn lọc cộng tác dựa trên người dùng, Độ đo tương đồng hàm ý thống kê.*

I. GIỚI THIỆU

Hệ tư vấn lọc cộng tác dựa trên người dùng là phiên bản đầu tiên của hệ tư vấn dựa trên lọc cộng tác. Nó được giới thiệu lần đầu tiên trong bài báo "GroupLens: an open architecture for collaborative filtering of netnews" vào năm 1994 cho hệ tư vấn GroupLens Usenet [12]. Đây là hệ tư vấn lọc cộng tác dựa trên người dùng đầu tiên hỗ trợ người đọc tìm ra các bài báo mà họ thích trong số lượng lớn các bài báo. Hệ thống tự động thu thập các giá xếp hạng của người dùng trong quá khứ để dự đoán sở thích của người dùng hiện tại. Với kiến trúc hoàn toàn mở, nên hệ thống có thể được phát triển mở rộng dễ dàng. Tiếp theo GroupLens Usenet, có hai hệ tư vấn khác cũng sử dụng phương pháp tư vấn lọc cộng tác dựa trên người dùng, đó là hệ tư vấn cho người dùng nghe nhạc Ringo [15] và hệ tư vấn cho người dùng xem phim BellCore [16]. Hệ tư vấn lọc cộng tác dựa trên người dùng là một giải thuật đơn giản làm sáng tỏ các tiền đề cốt lõi của phương pháp tư vấn lọc cộng tác. Đó là tìm ra người dùng trong quá khứ có hành vi tương đồng với người dùng hiện tại. Sau đó, sử dụng kết quả xếp hạng của họ cho các mặt hàng hay các mục dữ liệu để dự đoán sở thích của người dùng hiện tại. Như vậy, để có được danh sách các sản phẩm hay mục dữ liệu giới thiệu cho người dùng mới, hệ tư vấn lọc cộng tác dựa trên người dùng yêu cầu phải có hàm để tính sự tương đồng của hai người dùng và phương pháp tính độ lệch trung bình giá trị xếp hạng của những người dùng tương đồng với người dùng mới dựa trên ma trận xếp hạng của người dùng cho các sản phẩm hay các mục dữ liệu [9][8][11].

Trong bài viết này, dựa trên mô hình của hệ tư vấn lọc cộng tác truyền thống sử dụng các độ đo tương đồng như: Pearson, Cosine, Jaccard, chúng tôi đề xuất một tiếp cận mới cho hệ tư vấn lọc cộng tác dựa trên người dùng - Hệ tư vấn lọc cộng tác theo người dùng dựa trên độ đo hàm ý thống kê. Trong hệ thống này, chúng tôi xây dựng một độ đo tương đồng dựa trên độ đo chỉ số hàm ý thống kê (Implication index) để xác định sự tương đồng giữa hai người dùng gọi là độ đo tương đồng hàm ý thống kê để thay thế cho độ đo tương đồng trong hệ thống. Cụ thể hơn, giá trị tương đồng giữa hai người dùng được xác định dựa trên luật hàm ý thống kê theo khuynh hướng không đối xứng. Trong đó, các luật hàm ý thống kê được sinh ra dựa trên các phản ví dụ (counter-example) [14] và giá trị tương đồng được xác định dựa trên tần suất xuất hiện của các tham số n, n_A, n_B, n_{AB}, $n_{A\bar{B}}$ trong bảng phân phối xác suất 2x2. Mô hình tư vấn lọc cộng tác dựa trên người dùng sử dụng đo tương đồng hàm ý thống kê được chúng tôi xây dựng và triển khai thực nghiệm trên hai tập dữ liệu MovieLense [2] và MSWeb [7] đồng thời so sánh kết quả với mô hình tư vấn lọc cộng tác theo người dùng sử dụng các độ đo tương đồng truyền thống.

Bài viết này được tổ chức thành 6 phần. Phần 1 giới thiệu về hệ tư vấn lọc cộng tác dựa trên người dùng, các nghiên cứu liên quan và nêu vấn đề nghiên cứu. Phần 2 giới thiệu về mô hình hệ tư vấn lọc cộng tác dựa trên người dùng. Phần 3 trình bày cách xây dựng độ đo tương đồng của hai người dùng dựa trên độ đo hàm ý thống kê. Phần 4 mô tả các bước xây dựng hệ tư vấn theo người dùng dựa trên độ đo hàm ý thống kê. Phần 5 trình bày kết quả thực nghiệm của mô hình và so sánh kết quả với các mô hình khác. Phần cuối cùng tóm tắt một số kết quả quan trọng đã đạt được.

II. HỆ TƯ VẤN LỌC CỘNG TÁC DỰA TRÊN NGƯỜI DÙNG

Hệ tư vấn lọc cộng tác dựa trên người dùng là hệ thống tìm ra mục dữ liệu hay sản phẩm tương đồng để giới thiệu cho người dùng hiện tại dựa trên giá trị xếp hạng của các người dùng khác trong quá khứ. Những mục dữ liệu hay

sản phẩm có hạng cao nhất sẽ được dùng để tư vấn cho người dùng. Một cách hình thức, mô hình tư vấn lọc cộng tác dựa trên người dùng được mô tả như sau:

Gọi $U = \{u_1, u_2, \dots, u_m\}$ là tập m người dùng, $I = \{i_1, i_2, \dots, i_n\}$ là tập n sản phẩm hay mục dữ liệu, $R = r_{j,k}$ là ma trận xếp hạng của người dùng cho các sản phẩm hay mục dữ liệu với mỗi dòng biểu diễn cho một người dùng u_j $(1 \geq j \geq m)$, mỗi cột biểu diễn cho một sản phẩm hay mục dữ liệu i_k $(1 \geq k \geq n)$, $r_{j,k}$ là giá trị xếp hạng của người dùng u_j cho sản phẩm i_k và $u_a \in U$ là người dùng cần tư vấn.

2.1. Tính độ tương đồng giữa hai người dùng

Lựa chọn độ đo để tính độ tương đồng giữa hai người dùng là một khâu quan trọng trong việc thiết kế hệ tư vấn lọc cộng tác dựa trên người dùng. Bởi vì, nó ảnh hưởng trực tiếp đến kết quả tư vấn của hệ thống. Hiện tại, trong lĩnh vực nghiên cứu máy học, có nhiều độ đo được đề xuất cho mục đích này. Trong đó, Pearson correlation, Cosine similarity và Jaccard là ba độ đo được nhiều hệ tư vấn sử dụng.

Pearson correlation là độ đo tính sự tương đồng giữa hai người dùng dựa trên tương quan thống kê [8][9][13]. Độ tương đồng của hai người dùng *u* và *v* được xác định bằng công thức (1):

$$S(u,v) = \frac{\sum_{i \in I}(r_{v,i} - \bar{r}_v)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}} \quad (1)$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; *I* là tập các sản phẩm hay các mục dữ liệu được xếp hạng bởi cả hai người dùng; $r_{v,i}$ là giá trị xếp hạng của người dùng v cho sản phẩm i; $\bar{r}_v$ là giá trị xếp hạng trung bình của người dùng v; $r_{u,i}$ là giá trị xếp hạng của người dùng u cho sản phẩm i; $\bar{r}_u$ là giá trị xếp hạng trung bình của người dùng u;

Cosine similarity là độ đo tính sự tương đồng giữa hai người dùng dựa trên không gian vector đại số tuyến tính [8][9][13]. Các giá trị xếp hạng của từng người dùng trên m sản phẩm hay mục dữ liệu được biểu diễn bằng một vector m chiều. Độ tương đồng của hai người dùng *u* và *v* được xác định bằng khoảng cách Cosine giữa hai vector $\vec{r}_u$và vector $\vec{r}_v$ theo công thức (2):

$$S(u,v) = \cos(\vec{r}_u, \vec{r}_v) = \frac{\vec{r}_u \cdot \vec{r}_v}{\|\vec{r}_u\|_2 \, X \, \|\vec{r}_v\|_2} = \frac{\sum_{i=1}^{m} r_{u,i} r_{v,i}}{\sqrt{\sum_{i=1}^{m} r_{u,i}^2}\sqrt{\sum_{i=1}^{m} r_{v,i}^2}} \quad (2)$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; *m* là số chiều của vector (số sản phẩm); $r_{u,i}$ là giá trị xếp hạng của người dùng u cho sản phẩm i; $r_{v,i}$ là giá trị xếp hạng của người dùng v cho sản phẩm i;

Jaccard là độ đo xác định độ tương đồng giữa hai người dùng dựa trên các phép toán giữa hai tập hợp hữu hạn [11]. Đây là độ đo được đề xuất để xử lý vấn đề khi giá trị xếp hạng của người dùng cho một sản phẩm hay mục dữ liệu trong ma trận xếp hạng bị bỏ qua và được gán giá trị bằng 0. Trong độ đo này, tập các sản phẩm hay mục dữ liệu được người dùng *u* xếp hạng sẽ được xây dựng thành một tập hợp tương ứng trong hồ sơ người dùng *u*. Độ tương đồng của hai người dùng được tính bằng lực lượng phần tử của phép toán giao trên hai tập hợp (∩ - intersection) chia cho lực lượng phần tử của phép toán hợp trên hai tập hợp (∪ - union). Giá trị tương đồng giữa người dùng *u* và người dùng v được tính bằng công thức (3):

$$S(u,v) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; *A* là tập hợp biểu diễn các giá trị xếp hạng trên danh mục sản phẩm hay mục dữ liệu của người dùng u; *B* là tập hợp biểu diễn các giá trị xếp hạng trên danh mục sản phẩm hay mục dữ liệu của người dùng v.

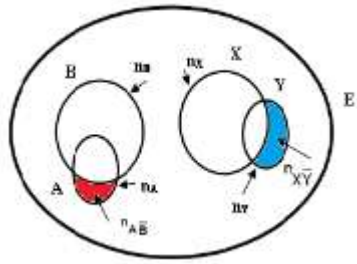
2.2. Sinh kết quả tư vấn

Để tính toán kết quả khuyến nghị cho người dùng *u*, bước đầu tiên, hệ tư vấn lọc cộng tác dựa trên người dùng sử dụng các độ đo tương đồng để tìm ra danh sách N người dùng tương đồng với người dùng *u*. Khi có danh sách N người dùng tương đồng, hệ thống sẽ kết hợp các giá trị xếp hạng của họ để sinh ra dự đoán sở thích của người dùng *u* đối với sản phẩm hay mục dữ liệu *i*. Thông thường, kết quả dự đoán được tính dựa trên trọng số trung bình của giá trị xếp hạng của N người dùng tương đồng được biểu diễn bằng công thức (4) [8][9]:

$$P(u,i) = \bar{r}_u + \frac{\sum_{u' \in N} S(u,u')(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |S(u,u')|} \quad (4)$$

III. ĐỘ ĐO TƯƠNG ĐỒNG HÀM Ý THỐNG KÊ CHO HAI NGƯỜI DÙNG

Phân tích hàm ý thống kê là phương pháp phân tích dữ liệu nghiên cứu mối quan hệ hàm ý giữa các biến hay thuộc tính dữ liệu, cho phép phát hiện các luật (rules A → B) không đối xứng theo dạng "nếu A sau đó gần như B" hoặc "xem xét đến mức độ nào mà B sẽ đáp ứng hàm ý của A" [14].



Hình 1. Mô hình biểu diễn luật hàm ý thống kê A → B

Mỗi luật dạng A → B được biểu diễn bởi một bảng dựa trên các khái niệm xác suất gọi là Bảng phân phối xác suất 2x2 để lưu trữ tần suất đếm được đáp ứng điều kiện cho trước. Từ Bảng phân phối xác suất này, các giá trị xác suất được tính dựa trên các tần suất xuất hiện n, n_A, n_B, n_{AB}, $n_{A\bar{B}}$ tương ứng như sau: $P(A) = \frac{n_A}{n}$, $P(B) = \frac{n_B}{n}$, $P(A \cap B) = \frac{n_{AB}}{n}$, $P(A \cap \bar{B}) = \frac{n_{A\bar{B}}}{n}$.

Bảng 1. Bảng phân phối xác suất 2x2 của luật A → B

	A → B		
	B	$\bar{B}$	
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$
	n_B	$n_{\bar{B}}$	n

Khi đó giá trị hấp dẫn (interestingness value) của độ đo hàm ý thống kê cho luật A → B được xác định bởi hàm dựa các bản số luật có trong bảng phân phối xác suất 2x2 có dạng theo công thức (5) [14]:

$$q(A \rightarrow \bar{B}) = \frac{n_{A\bar{B}} - \frac{n_A(n-n_B)}{n}}{\sqrt{\frac{n_A(n-n_B)}{n}}} \quad (5)$$

Từ công thức tính giá trị hấp dẫn của độ đo chỉ số hàm ý thống kê, khoảng cách hay sự khác nhau giữa hai người dùng u, v được xác bằng giá trị trung bình của các chỉ số hàm ý dựa trên các luật hàm ý thống kê mà vế trái là tập các sản phẩm hay mục dữ liệu nằm trong hồ sơ người dùng u và vế phải là tập các sản phẩm hay mục dữ liệu không nằm trong hồ sơ người dùng v theo công thức (6):

$$D_q(u,v) = \left| \frac{\sum_{x \in I_u, y \notin I_v} q(x \rightarrow y)}{M} \right| \quad (6)$$

Với $D_q(u,v)$ là giá trị khoảng cách giữa người dùng u và người dùng v; I_u là tập các sản phẩm hay mục dữ liệu được người dùng u xếp hạng; I_v là tập các sản phẩm hay mục dữ liệu được người dùng v xếp hạng; x là tập con của tập I_u, y là tập con bù của tập I_v và x, y là hai tập phần tử rời nhau ($x \cap y = \emptyset$); M là tổng số luật hàm ý thống kê được sinh ra.

Dựa trên công thức tính khoảng cách giữa hai người dùng, chúng tôi đề xuất một độ đo tương đồng dùng để đo giá trị tương đồng giữa hai người dùng gọi là độ đo tương đồng hàm ý thống kê và được tính bằng công thức (7):

$$S_q(u,v) = 1 - D_q(u,v) \quad (7)$$

IV. HỆ TƯ VẤN LỌC CỘNG TÁC THEO NGƯỜI DÙNG DỰA TRÊN ĐỘ ĐO TƯƠNG ĐỒNG HÀM Ý THỐNG KÊ

Dựa trên các bước xây dựng hệ tư vấn lọc công tác dựa trên người dùng [8][9][11] và độ đo xác định độ tương đồng giữa hai người dùng dựa trên độ đo hàm ý thống kê đã trình bày ở phần trên, chúng tôi đề xuất các bước xây dựng hệ tư vấn lọc cộng tác theo người dùng dựa trên độ đo tương đồng hàm ý thống kê như sau:

4.1. Tìm hiểu bài toán thực tế

Hệ tư vấn là công cụ hỗ trợ người dùng lựa chọn các sản phẩm và dịch vụ thông qua mạng Internet. Vì thế, việc nắm bắt được các kỳ vọng của người dùng, những gì mà họ cần cung cấp từ hệ thống, là một bước rất quan trọng trong quy trình xây dựng hệ tư vấn. Nó chính là cơ sở cho việc lựa chọn và xử lý dữ liệu và cùng là điều kiện cần để đưa hệ thống vào ứng dụng thực tế.

4.2. Lựa chọn và xử lý dữ liệu

Việc lựa chọn và xử lý dữ liệu cũng được đánh giá là bước quan trọng trong tiến trình xây dựng hệ thống tư vấn. Bởi do đặc thù của hệ tư vấn là kết quả tư vấn được sinh ra dựa trên sự tính toán từ dữ liệu cũ. Cụ thể hơn, hệ tư vấn đưa ra các khuyến nghị cho người dùng trong tương lai bằng cách áp dụng các phép toán thống kê, các thuật toán

máy học trên các tập dữ liệu chuyên ngành được thu thập trong quá khứ. Vì vậy, lựa chọn đúng và xử lý tốt dữ liệu sẽ góp phần làm tăng độ chính xác kết quả khuyến nghị của hệ thống.

4.3. Xây dựng giải thuật tư vấn

Trong giải thuật này, khi có một người dùng mới cần tư vấn, hệ thống sẽ sử dụng độ đo tương đồng giữa hai người dùng dựa trên độ đo tương đồng hàm ý thống kê để tìm ra danh sách các người dùng tương đồng với người dùng mới. Sau đó, danh sách các sản phẩm hay mục dữ liệu được xếp hạng cao để giới thiệu cho người dùng mới được tính toán dựa trên các giá trị xếp hạng của các người dùng tương đồng. Giải thuật hệ tư vấn được thực hiện theo các bước sau:

Input: Ma trận xếp hạng của người dùng, Người dùng cần tư vấn u_a.

Output: Danh sách các sản phẩm hay các mục dữ liệu có giá trị xếp hạng cao.

Begin

Bước 1: Cập nhật hoặc xây dựng hồ sơ người dùng mới u_a.

Bước 2: Tìm danh sách người dùng tương đồng:

- Chọn hệ số k để xác định danh sách k người dùng tương đồng với người dùng mới.
- Xác định danh sách người dùng tương đồng dựa trên độ đo tương đồng hàm ý thống kê.

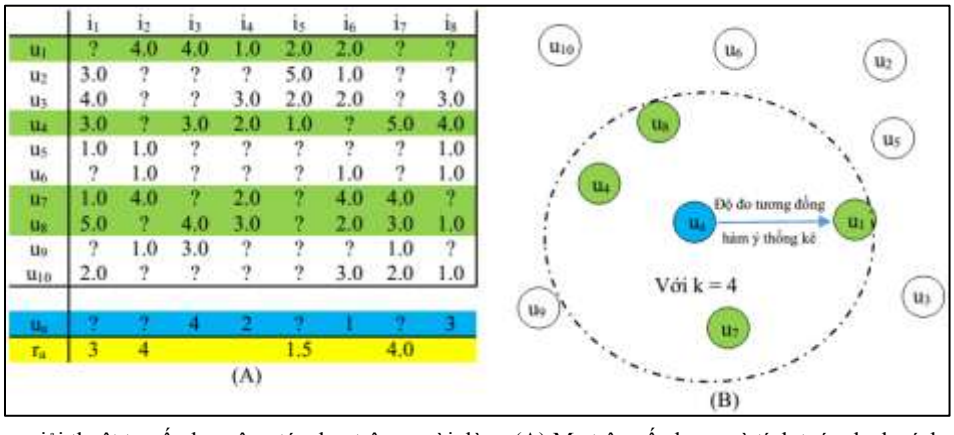
Bước 3: Tìm danh sách các sản phẩm hay mục dữ liệu được xếp hạng cao nhất bởi k người dùng tương đồng:

- Tính giá trị xếp hạng trung bình cho từng sản phẩm hay mục dữ liệu.
- Sắp xếp các sản phẩm hay mục dữ liệu dựa trên giá trị trung bình xếp hạng.

Bước 4: Chọn N sản phẩm hay mục dữ liệu có trị trung bình xếp hạng cao nhất làm kết quả tư vấn.

End.

Để thấy rõ hơn các bước thực hiện của giải thuật, chúng tôi biểu diễn giải thuật tư vấn bằng ví dụ cụ thể, được trình bày trong hình 2. Trong hình này, giả sử hệ thống tư vấn cho người dùng lựa chọn 8 sản phẩm (từ i_1 đến i_8) và hiện tại hệ thống đã có 10 người dùng (từ u_1 đến u_{10}) xếp hạng cho các sản phẩm. Các sản phẩm được xếp hạng từ 1 đến 5 ("1"- đánh giá thấp nhất; "5" – đánh giá cao nhất; "?" – người dùng không đánh giá cho sản phẩm). Hệ thống được yêu cầu tư vấn các sản phẩm cho người dùng mới u_a, với hệ số k được cho bằng 4 (tính trên 4 người dùng tương đồng). Từ yêu cầu này, hệ thống đã tìm ra danh sách 4 người dùng tương đồng với u_a là u_1, u_4, u_7 và u_8 để đoán ra các giá trị xếp hạng mà người dùng u_a bỏ qua và xác định danh sách các sản phẩm mà hệ thống đoán là u_a sẽ thích.



	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	1.0	2.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	4.0	?	?	3.0	2.0	2.0	?	3.0
u_4	3.0	?	3.0	2.0	1.0	?	5.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	?	1.0	?	1.0
u_7	1.0	4.0	?	2.0	?	4.0	4.0	?
u_8	5.0	?	4.0	3.0	?	2.0	3.0	1.0
u_9	?	1.0	3.0	?	?	?	1.0	?
u_{10}	2.0	?	?	?	?	3.0	2.0	1.0
u_a	?	?	4	2	?	1	?	3
r_a	3	4			1.5		4.0	

Hình 2. Minh họa giải thuật tư vấn lọc cộng tác dựa trên người dùng (A) Ma trận xếp hạng và tính toán danh sách các sản phẩm dự đoán cho người dùng u_a; (B) Xác định người dùng tương đồng với người dùng mới

4.4. Đánh giá mô hình tư vấn

Đánh giá độ chính xác của mô hình tư vấn là một khâu quan trọng trong quy trình xây dựng hệ tư vấn [5][4][3]. Nó giúp cho người thiết kế mô hình lựa chọn mô hình, kiểm tra độ chính xác của mô hình trước khi đưa mô hình vào ứng dụng thực tế. Để đánh giá mô hình tư vấn lọc công tác dựa trên người dùng, người xây dựng hệ thống cần thực hiện qua 2 bước sau:

4.4.1. Chuẩn bị dữ liệu cho đánh giá

Để đánh giá mô hình tư vấn, bước đầu tiên là chuẩn bị dữ liệu, trong bước này tập dữ liệu thực nghiệm được chia làm hai tập con: tập dữ liệu huấn luyện (Training set) và tập dữ liệu kiểm tra (Testing set) [10][11]. Hiện tại, có nhiều phương pháp để chia tập dữ liệu cho việc đánh giá mô hình tư vấn như:

Splitting: là phương pháp đầu tiên để xây dựng tập huấn luyện và tập kiểm tra bằng cách cắt tập dữ liệu thực nghiệm thành 2 phần. Với phương pháp này, người thiết kế mô hình cần quyết định tỷ lệ phần trăm cho tập huấn luyện và tập kiểm tra. Ví dụ, tập huấn luyện chiếm 80 phần trăm và tập kiểm tra chiếm 20 phần trăm còn lại.

Bootstrap sampling: là phương pháp xây dựng tập huấn luyện và tập kiểm tra bằng cách cắt tập dữ liệu thực nghiệm thành 2 phần. Tuy nhiên, việc cắt này được thực hiện ngẫu nhiên nhiều lần nhằm mục đích để một người dùng có thể là thành viên của tập huấn luyện ở lần cắt này nhưng là thành viên của tập kiểm tra trong các lần cắt tiếp theo. Điều này có thể khắc phục được nhược điểm không đồng đều của tập dữ liệu thực nghiệm đồng thời tăng tính tối ưu trên các tập dữ liệu có kích thước nhỏ.

K-fold cross-validation: là phương pháp xây dựng tập huấn luyện và tập kiểm tra bằng cách cắt tập dữ liệu thực nghiệm thành k tập con có kích cỡ giống nhau (gọi là k-fold). Sau đó, thực hiện k lần đánh giá, với mỗi lần đánh giá sử dụng một tập con làm tập kiểm tra và k-1 tập con còn lại dùng làm tập huấn luyện. Kết quả đánh giá được tính từ kết quả k lần kiểm tra bằng phép tính trung bình. Phương pháp này đảm bảo rằng mọi người dùng ít nhất một lần xuất hiện trong tập kiểm tra. Vì thế, nó có độ chính xác cao nhất trong ba phương pháp được nêu ở trên. Tuy nhiên, nó mất nhiều chi phí cho việc tính toán so với hai phương pháp còn lại.

4.4.2. Đánh giá mô hình tư vấn

Có hai phương pháp để đánh giá mô hình tư vấn: đánh giá dựa trên các xếp hạng (Evaluation the ratings) và đánh giá dựa trên các khuyến nghị (Evaluation the recommendations) [10]. Phương pháp đầu đánh giá các xếp hạng được sinh ra bởi mô hình. Phương pháp còn lại đánh giá trực tiếp trên các khuyến nghị của mô hình.

Đánh giá dựa trên các xếp hạng: phương pháp này đánh giá độ chính xác của mô hình bằng cách so sánh giá trị xếp hạng dự đoán với giá trị thực hay chính xác hơn là tìm ra giá trị trung bình lỗi dựa vào ba đại lượng RMSE, MSE và MAE [5][10]. Mô hình được đánh giá là tốt khi các đại lượng này có giá trị thấp.

Root mean square error (RMSE): Độ lệch chuẩn giữa giá trị thực và giá trị xếp hạng.

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in\kappa}(r_{ij}-\hat{r}_{ij})^2}{|\kappa|}}$$

Mean squared error (MSE): Trung bình của bình phương độ lệch giữa giá trị thực và giá trị xếp hạng dự đoán. Nó chính là bình phương của độ lệch chuẩn.

$$MSE = \frac{\sum_{(i,j)\in\kappa}(r_{ij}-\hat{r}_{ij})^2}{|\kappa|}$$

Mean absolute error (MAE): Trung bình trị tuyệt đối của độ lệch giữa giá trị thực và giá trị xếp hạng dự đoán.

$$MAE = \frac{1}{|\kappa|}\sum_{(i,j)\in\kappa}|r_{ij} - \hat{r}_{ij}|$$

Với κ là tập tất cả các xếp hạng của người dùng cho sản phẩm hay mục dữ liệu; r_{ij} giá trị xếp hạng thực của người dùng i cho sản phẩm hay mục dữ liệu j; $\hat{r}_{ij}$ là giá trị xếp hạng dự đoán của người dùng i cho sản phẩm hay mục dữ liệu j.

Đánh giá dựa trên các khuyến nghị: phương pháp này đánh giá độ chính xác của mô hình bằng cách so sánh các khuyến nghị của mô hình đưa ra với các lựa chọn mua hay không mua của người dùng. Phương pháp này sử dụng ma trận hỗn độn 2x2 (Confusion matrix) để tính độ chính xác (Precision), độ bao phủ (Recall) và trung bình điều hòa giữa độ chính xác và độ bao phủ (F-measure) [5][10]. Mô hình được đánh giá là tốt khi ba chỉ số trên có giá trị cao.

Bảng 2. Ma trận hỗn độn

Lựa chọn của người dùng	Khuyến nghị của mô hình	
	Giới thiệu	Không giới thiệu
Mua	**TP**	**FN**
Không mua	**FP**	**TN**

Trong đó:

TP: Những sản phẩm được mô hình khuyến nghị đã được mua.

FP: Những sản phẩm được mô hình khuyến nghị không được mua.

FN: Những sản phẩm không được mô hình khuyến nghị đã được mua.

TN: Những sản phẩm không được mô hình khuyến nghị không được mua.

$$\text{Độ } chính\ xác \text{ (Precision)} = \frac{Số\ sản\ phẩm\ giới\ thiệu\ chính\ xác}{Tổng\ số\ sản\ phẩm \text{ được } giới\ thiệu} = \frac{TP}{TP+FP}$$

$$\text{Độ } bao\ phủ \text{ (Recall)} = \frac{Số\ sản\ phẩm\ giới\ thiệu\ chính\ xác}{Tổng\ số\ sản\ phẩm \text{ được } mua} = \frac{TP}{TP+FN}$$

$$Trung\ bình \text{ điều } hòa\ (F-measure) = \frac{2*\text{Độ } chính\ xác * \text{Độ } bao\ phủ}{\text{Độ } chính\ xác + \text{Độ } bao\ phủ}$$

V. THỰC NGHIỆM

5.1. Dữ liệu sử dụng

Trong phần thực nghiệm này, chúng tôi sử dụng hai tập dữ liệu khác nhau để chạy mô hình trên hai kịch bản khác nhau. Kịch bản 1, chúng tôi sử dụng tập dữ MovieLense [2] của dự án nghiêm cứu GroupLens tại Trường Đại học

Minnesota (the University of Minnesota) vào năm 1997. Kịch bản 2, chúng tôi sử dụng tập dữ liệu MSWeb của Microsoft công bố vào năm 1998 [7].

Trong kịch bản 1, chúng tôi tiến hành thực nghiệm trên tập dữ liệu MovieLense. Đây là tập dữ liệu được thu thập từ kết quả xếp hạng của 943 người dùng cho 1.664 bộ phim (99.392 kết quả xếp hạng từ 1 đến 5) thông qua trang web MovieLense (movielens.umn.edu) trong thời gian 7 tháng (từ 19/9/1997 đến 22/4/1998). Tập dữ liệu này được tổ chức theo định dạng ma trận gồm 943 hàng, 1.664 cột và 1.569.152 ô chứa giá trị xếp hạng. Tuy nhiên, không phải mỗi người dùng đều xem tất cả các phim. Vì thế, trong ma trận xếp hạng chỉ có 99.392 lượt xếp hạng của người dùng cho danh mục các bộ phim.

Trong kịch bản 2, chúng tôi tiến hành thực nghiệm trên tập dữ liệu MSWeb. Đây là tập dữ liệu về người dùng Microsoft truy cập các trang web trong thời gian một tuần trong tháng 2 năm 1998 được lấy mẫu và xử lý từ file log của địa chỉ www.microsoft.com. Tập dữ liệu này bao gồm 38.000 người dùng nặc danh truy cập trên 285 địa chỉ web gốc được xử lý và tổ chức thành ma trận nhị phân với 32.710 hàng, 285 cột và 98.653 giá trị xếp hạng.

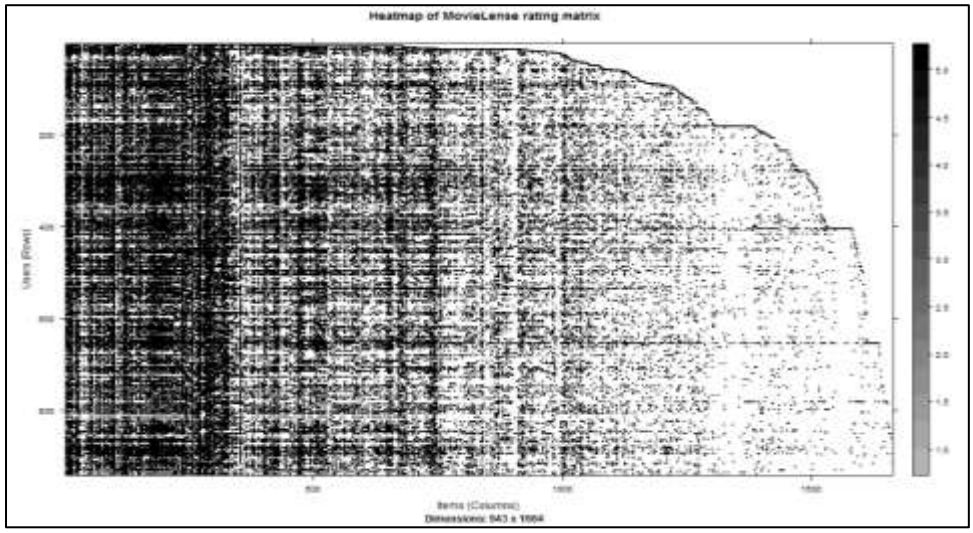
5.2. Công cụ thực hiện (ARQAT)

Để triển khai hai kịch bản thực nghiệm, chúng tôi sử dụng công cụ ARQAT được triển khai trên ngôn ngữ R. Đây là gói công cụ được nhóm chúng tôi phát triển từ nền tảng của công cụ ARQAT phát triển trên Java [6]. Công cụ này gồm các chức năng: xử lý dữ liệu, hàm sinh luật hàm ý thống kê, hàm đếm các tham số $n, n_A, n_B, n_{A\bar{B}}$, các hàm tính độ đo hấp dẫn khách quan cho luật hàm ý dựa trên 4 tham số hàm thống kê n, $n_A, n_B, n_{A\bar{B}}$, chức năng tính độ tương đồng của hai đối tượng dựa trên độ đo chỉ số hàm ý, các chức năng xây dựng và đánh giá mô hình của hệ tư vấn [10].

5.3. Kịch bản 1

5.3.1. Lựa chọn và xử lý dữ liệu

Tập dữ liệu MovieLense được lưu trữ dạng ma trận xếp hạng số thực (giá trị từ 0 đến 5) gồm 943 dòng, 1.664 cột với 1.569.152 ô chứa giá trị xếp hạng. Trong đó, hơn 93 phần trăm giá trị xếp hạng có giá trị bằng 0 và 7 phần trăm còn lại có giá trị xếp hạng có giá trị từ 1 đến 5 (Giá trị hạng 0 là 1.469.760; Giá trị hạng 1 là 6.059; Giá trị hạng 2 là 11.307; Giá trị hạng 3 là 27.002; Giá trị hạng 4 là 33.947; Giá trị hạng 5 là 21.077). Như vậy, toàn tập dữ liệu MovieLense chỉ có 99.392 giá trị xếp hạng của người dùng thật sự xếp hạng cho các bộ phim. Trong đó, đa số giá trị xếp hạng nằm trong khoảng từ 3 đến 5 và 4 là giá trị xếp hạng chiếm số lượng cao nhất. Để thấy rõ phân bố giá trị xếp hạng của tập dữ liệu MovieLense chúng tôi sử dụng biểu đồ nhiệt để đại diện cho các giá trị xếp hạng của người dùng được trình bày trong hình 3.



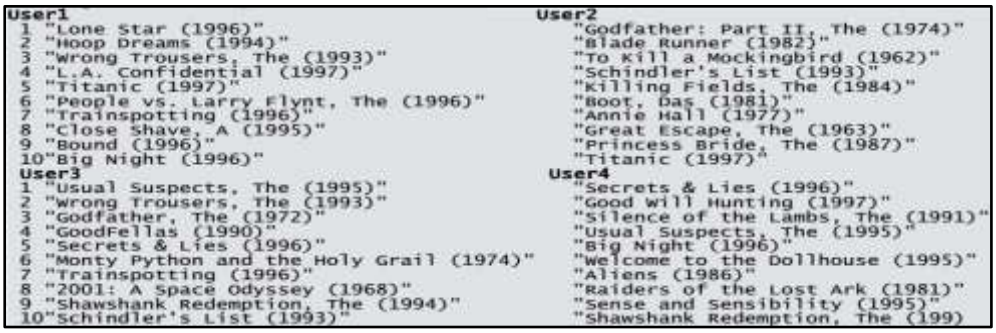
Hình 3. Biểu đồ nhiệt trình bày phân bố giá trị xếp hạng của người dùng trên tập dữ liệu MovieLense

Từ biểu đồ nhiệt, chúng tôi thấy rằng có số phim chỉ được xếp hạng bởi một vài người dùng và có một số người dùng chỉ xếp hạng cho một vài phim. Nếu sử dụng các trường hợp này để huấn luyện mô hình có thể dẫn đến các sai lệch do thiếu thông tin. Vì vậy, chúng tôi quyết định chỉ chọn các người dùng có xếp hạng ít nhất cho 50 phim và các bộ phim phải được ít nhất 100 người dùng xếp hạng để xây dựng tập dữ thực nghiệm cho mô hình. Khi đó, ma trận dữ liệu thực nghiệm chỉ còn 560 dòng, 332 cột và 55.298 giá trị xếp hạng. Trong đó, chúng tôi chia tập dữ liệu thành hai tập con: Tập huấn luyện chiếm tỷ lệ 80 phần trăm và Tập kiểm tra chiếm tỷ lệ 20 phần trăm còn lại.

5.3.2. Kết quả của mô hình

Với mục tiêu kiểm tra độ chính xác của mô hình trên ma trận xếp hạng dạng số thực, chúng tôi tiến hành xây dựng mô hình tư vấn theo người dùng dựa trên độ đo tương đồng hàm ý thống kê trên tập dữ liệu huấn luyện với 449 người dùng và kiểm tra kết quả của mô hình tập dữ liệu kiểm tra với 111 người dùng. Trong đó, số người dùng tương đồng được xác định là k=25. Kết quả tư vấn của mô hình được xuất ra theo định dạng ma trận với cấu trúc 10 x 111

(mỗi cột là một người dùng, mỗi ô là một phim được chọn để giới thiệu cho người dùng ở cột tương ứng). Hình 4 trình bày kết quả tư vấn cho 4 người dùng đầu tiên, với mỗi người dùng chọn 10 phim được xếp hàng cao nhất.



Hình 4. Trình bày kết quả tư vấn của 4 người dùng đầu tiên

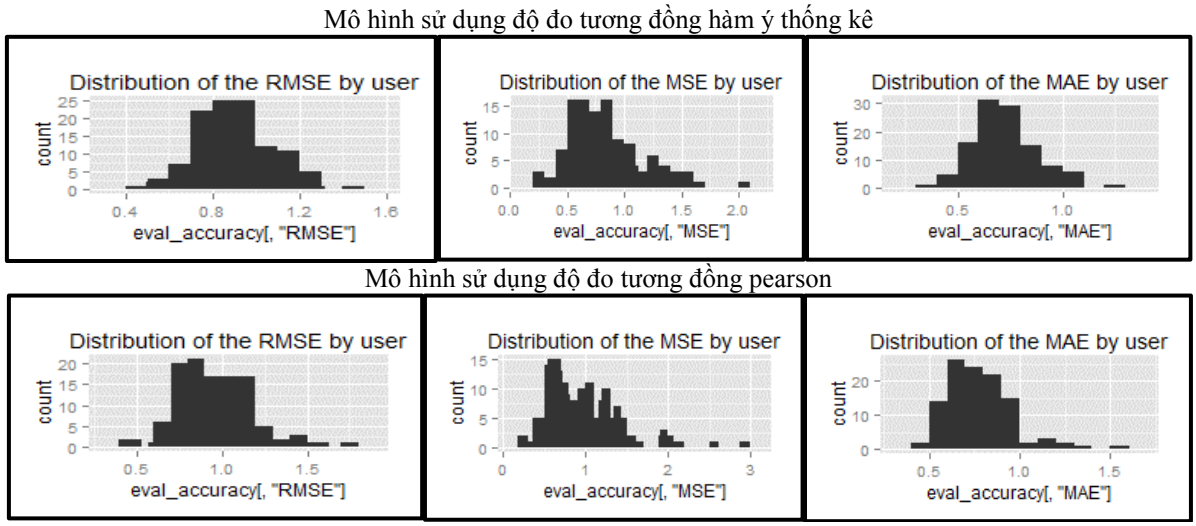
Dựa trên ma trận kết quả tư vấn, chúng tôi thống kê số lần được giới thiệu của từng bộ phim. Qua kết quả thống kê, chúng ta thấy rằng phần lớn phim chỉ được giới thiệu dưới 5 lần. Trong đó, có đến 38 phim chỉ được giới thiệu duy nhất 1 lần, 26 phim được giới thiệu 2 lần. Ngược lại, số phim được giới thiệu từ 5 lần trở lên chiếm số lượng tương đối nhỏ. Đa số đều có số lượng dưới 5. Trong đó có 2 phim được giới thiệu đến 41 lần.

5.3.3. Đánh giá mô hình

Trong phần này, chúng tôi tính các thông số lỗi (RMSE, MSE, MAE) cho từng người dùng và cho mô hình dựa trên dữ liệu được xây dựng bằng phương pháp k-fold (k=5). Đối với thông số lỗi của từng người dùng, chúng tôi trình bày phân bố của từng thông số lỗi bằng dạng biểu đồ đồng thời so sánh các thông số này với các thông số lỗi của mô hình sử dụng độ đo tương đồng Pearson (hình 5). Qua biểu đồ ta thấy rằng, số lượng người dùng phân bố trên các thông số lỗi của mô hình sử dụng độ đo tương đồng hàm ý thống kê có giá trị cao hơn so với số lượng người dùng phân bố trên các thông số lỗi của mô hình sử dụng độ đo tương đồng Pearson. Đối với thông số lỗi của toàn mô hình, chúng tôi tiếp tục so sánh với mô hình sử dụng độ đo tương đồng Pearson được trình bày trong bảng 3. Qua kết quả so sánh, chúng tôi thấy rằng mô hình của chúng đề xuất có các thông số lỗi thấp hơn mô hình sử dụng độ đo tương đồng Pearson trên tập dữ liệu MovieLense.

Bảng 3. Trình bày so sánh các thông số lỗi của hai mô hình

	RMSE	MSE	MAE
Mô hình sử dụng độ đo tương đồng hàm ý thống kê	0.8961562	0.8030960	0.7077939
Mô hình sử dụng độ đo Pearson	0.9796664	0.9597462	0.7704055



Hình 5. Trình bày so sánh các thông số lỗi của từng người dùng trên hai mô hình

5.4. Kịch bản 2

5.4.1. Lựa chọn và xử lý dữ liệu

Ma trận dữ liệu nhị nhân MSWeb có kích thước đương đối lớn với 32.710 dòng, 285 cột, 98.635 giá trị xếp hạng. Tuy nhiên, qua khảo sát chung tôi thấy rằng có khá nhiều người dùng chỉ truy cập một vài trang web và khá nhiều trang web chỉ được truy cập bởi một vài người dùng. Để tăng độ tin cậy của kết quả khuyến nghị của mô hình, chúng tôi tiến hành xây dựng tập dữ liệu cho mô hình theo điều kiện chỉ chọn những người dùng truy cập ít nhất 10 địa chỉ web và những trang web được truy cập ít nhất 50 người dùng. Sau khi thực hiện các thao tác chọn lọc, chúng tôi đã có ma trận dữ liệu cho thực nghiệm có kích thước 796x135. Tương tự như trong kịch bản 1, ma trận dữ liệu thực

nghiệm được chia làm hai tập con: Tập dữ liệu huấn luyện có kích thước 626x135 (chiếm 80%), Tập dữ liệu kiểm tra có kích thước 170x135 (chiếm 20%).

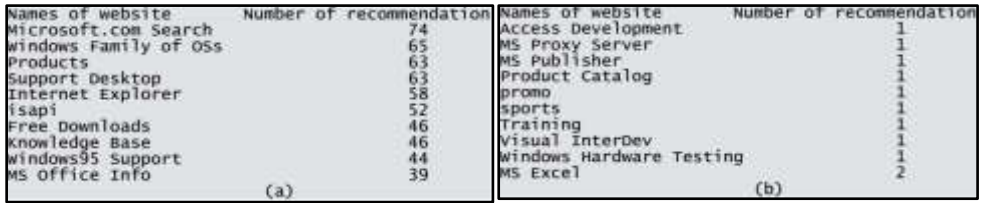
5.4.2. Kết quả của mô hình

Với mục tiêu kiểm tra độ chính xác của mô hình trên ma trận xếp hạng nhị phân, chúng tôi tiến xây dựng mô hình tư vấn theo người dùng dựa trên độ đo tương đồng hàm ý thống kê trên tập dữ liệu đã được xử lý ở trên. Trong lần thực nghiệm này, chúng tôi chọn số người dùng tương đồng k=30 và mỗi người dùng được giới thiệu 6 trang web mà mô hình dự đoán là họ yêu thích. Kết quả tư vấn cho 6 người dùng đầu tiên được trình bày trong hình 6.

User1	User2
1 "Free Downloads"	"Support Desktop"
2 "Support Desktop"	"Knowledge Base"
3 "Microsoft.com Search"	"Internet Site Construction for Developers"
4 "Windows Family of OSs"	"SiteBuilder Network Membership"
5 "Products"	"Windows NT Server"
6 "Developer Workshop"	"Developer Workshop"
User3	**User4**
1 "Developer Workshop"	"isapi"
2 "Microsoft.com Search"	"Developer Network"
3 "Developer Network"	"MS Office Development"
4 "Windows NT Server"	"OutLook"
5 "Products"	"ActiveX Technology Development"
6 "ActiveX Technology Development"	"MS Office"
User5	**User6**
1 "Microsoft.com Search"	"Microsoft.com Search"
2 "Free Downloads"	"Windows Family of OSs"
3 "Windows Family of OSs"	"Internet Explorer"
4 "MS Office"	"Windows95 Support"
5 "Internet Site Construction for Developers"	"IT Technical Information"
6 "SiteBuilder Network Membership"	"MS Word"

Hình 6. Trình bày kết quả khuyến nghị trên tập MSWeb cho 6 users đầu tiên

Từ ma trận kết quả tư vấn, chúng tôi trích chọn 10 trang web được mô hình chọn để giới thiệu đến người dùng nhiều nhất (hình 7a). Trong đó, đứng đầu danh sách là trang tìm kiếm của Microsoft (Microsoft.com Search) với 74 lần giới thiệu đến người dùng. Ngược lại với danh sách 10 trang web đứng đầu, chúng tôi cũng trích ra danh sách 10 trang web có số lần giới thiệu ít nhất (hình 7b). Hầu hết các trang này chỉ được giới thiệu duy nhất 1 lần ngoại trừ trang Microsoft Excel được giới chọn để giới thiệu 2 lần.



Names of website	Number of recommendation
Microsoft.com Search	74
Windows Family of OSs	65
Products	63
Support Desktop	63
Internet Explorer	58
isapi	52
Free Downloads	46
Knowledge Base	46
Windows95 Support	44
MS Office Info	39

(a)

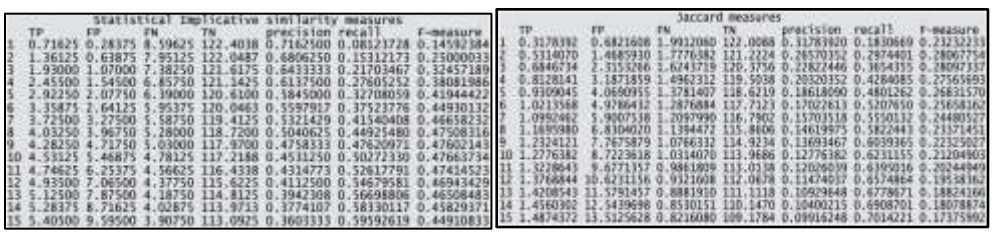
Names of website	Number of recommendation
Access Development	1
MS Proxy Server	1
MS Publisher	1
Product Catalog	1
promo	1
sports	1
Training	1
Visual InterDev	1
Windows Hardware Testing	1
MS Excel	2

(b)

Hình 7. Danh sách 10 trang web được giới thiệu nhiều nhất và 10 trang được giới thiệu ít nhất

5.4.3. Đánh giá mô hình

Do tập dữ liệu MSWeb là tập dữ liệu nhị phân, nên mô hình chỉ được đánh giá dựa trên các khuyến nghị. Trong đánh giá này, chúng tôi tiếp tục sử dụng phương pháp chia dữ liệu k-fold để xây dựng bộ dữ liệu đánh giá với k=5. Để khảo sát độ chính xác của mô hình, chúng tôi kiểm tra mô hình với số lượng trang web được giới thiệu đến người dùng tăng dần (từ 1 đến 15). Kết quả đánh giá trung bình của 5 k-fold trên mô hình sử dụng độ đo tương đồng hàm ý thống kê và mô hình sử dụng độ đo tương đồng Jaccard được trình bày trong hình 8. Trong hình này, chúng tôi thấy rằng chỉ số precision luôn giảm khi số trang web được giới thiệu tăng trên cả hai độ đo tương đồng. Ngược lại với precision, chỉ số recall luôn tăng trên cả hai độ đo tương đồng khi số trang web được giới thiệu tăng. Khác với hai chỉ số trên, chỉ số F-measure đạt giá trị cao nhất khi số trang web được giới thiệu bằng 10 trên mô hình sử dụng độ đo tương đồng hàm ý thống kê và bằng 3 trên mô hình sử dụng độ đo tương đồng Jaccard. Điều này cho thấy rằng mô hình sử dụng độ đo tương đồng hàm ý thống kê cho kết quả tốt nhất khi mỗi người dùng được giới thiệu 10 trang web và mô hình sử dụng độ đo tương đồng Jaccard cho kết quả tốt nhất khi mỗi người dùng được giới thiệu 3 trang web.



Statistical Implicative similarity measures

	TP	FP	FN	TN	precision	recall	F-measure
1	0.71825	0.28375	8.59625	122.4038	0.7162500	0.08121728	0.14592384
2	1.36125	0.63875	7.95125	122.0487	0.6806250	0.15312173	0.25000033
3	1.93000	1.07000	7.38250	121.6175	0.6433333	0.21703467	0.32457189
4	2.45500	1.54500	6.85750	121.1425	0.6137500	0.27605252	0.38081886
5	2.92250	2.07750	6.39000	120.6100	0.5845000	0.32708059	0.41944422
6	3.35875	2.64125	5.95375	120.0463	0.5597917	0.37523776	0.44930132
7	3.72500	3.27500	5.58750	119.4125	0.5321429	0.41540408	0.46658232
8	4.03250	3.96750	5.28000	118.7200	0.5040625	0.44925480	0.47508316
9	4.28250	4.71750	5.03000	117.9700	0.4758333	0.47620971	0.47602143
10	4.53125	5.46875	4.78125	117.2188	0.4531250	0.50272330	0.47663734
11	4.74625	6.25375	4.56625	116.4338	0.4314773	0.52617791	0.47414523
12	4.93500	7.06500	4.37750	115.6225	0.4112500	0.54679581	0.46943429
13	5.12500	7.87500	4.18750	114.8125	0.3942308	0.56698806	0.46550443
14	5.28375	8.71625	4.02875	113.9713	0.3774107	0.58130117	0.45829371
15	5.40500	9.59500	3.90750	113.0925	0.3603333	0.59592619	0.44910833

Jaccard measures

	TP	FP	FN	TN	precision	recall	F-measure
1	0.3178392	0.6821608	1.9912080	122.0088	0.31783920	0.1830669	0.23232233
2	0.5314070	1.4685930	1.7776382	121.2224	0.26570352	0.2974401	0.28067754
3	0.6846734	2.3153266	1.6241719	120.3756	0.22822446	0.3854155	0.28047337
4	0.8128141	3.1871859	1.4962312	119.5038	0.20320352	0.4284085	0.27565693
5	0.9309045	4.0690955	1.3781407	118.6219	0.18618090	0.4801262	0.26831570
6	1.0213568	4.9786432	1.2876884	117.7123	0.17022613	0.5207650	0.25658162
7	1.0992462	5.9007538	1.2097990	116.7902	0.15703518	0.5550132	0.24480527
8	1.1695980	6.8304020	1.1394472	115.8606	0.14619975	0.5822443	0.23371451
9	1.2324121	7.7675879	1.0766332	114.9234	0.13693467	0.6039365	0.22325027
10	1.2776382	8.7223618	1.0314070	113.9686	0.12776382	0.6231155	0.21204803
11	1.3228643	9.6771357	0.9861909	113.0138	0.12026039	0.6399016	0.20244949
12	1.3768844	10.6231156	0.9121608	112.0678	0.11474037	0.6574864	0.19538162
13	1.4208543	11.5791457	0.8881910	111.1118	0.10929648	0.6778671	0.18824166
14	1.4560302	12.5439698	0.8530151	110.1470	0.10400215	0.6908701	0.18078874
15	1.4874372	13.5125628	0.8216080	109.1784	0.09916248	0.7014221	0.17375992

Hình 8. So sánh kết quả đánh giá trung bình của 5 k-fold khi số trang web được giới thiệu tăng dần từ 1 đến 15

VI. KẾT LUẬN

Trong bài viết này, chúng tôi đã xây dựng mô hình hệ tư vấn lọc cộng tác dựa trên người dùng bằng cách đề xuất một độ đo tương đồng dựa trên độ đo chỉ số hàm ý thống kê để xác định sự tương đồng của hai người dùng. Giống như các hệ tư vấn lọc cộng tác dựa trên người dùng khác, mô hình của chúng tôi vẫn đảm bảo các bước chính như: xử lý dữ liệu, xây dựng ma trận xếp hạng, tính độ tương đồng giữa hai người dùng, xác định danh sách sản phẩm được người dùng tương đồng xếp hạng cao để đưa ra kết quả khuyến nghị và đánh giá tính chính xác của mô hình. Tuy nhiên, điểm mới trong mô hình này là trong bước xác định danh sách người dùng tương đồng, thay vì sử dụng các độ đo tương đồng quen thuộc như Pearson correlation, Cosine similarity, Jaccard để xác định sự tương đồng giữa hai

người dùng, chúng tôi sử dụng độ đo tương đồng hàm ý thống kê. Qua thực nghiệm, chúng tôi thấy rằng của mô hình chúng tôi cho kết quả khuyến nghị khá chính xác trên tập dữ liệu số thực và tập dữ liệu nhị phân. Đối với tập dữ liệu số thực MovieLense, các chỉ số lỗi (RMSE, MSE, MAE) của mô hình có giá trị thấp hơn so với kết quả đánh giá của mô hình sử dụng độ đo tương đồng Pearson. Đối với tập dữ liệu nhị phân MSWeb, các chỉ số đánh giá độ chính xác Precision, Recall, F-measure của mô hình có giá trị vượt trội so với các chỉ số này trong mô hình sử dụng độ đo tương đồng Jaccard. Kết quả thực nghiệm này cho thấy rằng mô hình hệ tư vấn lọc cộng tác theo người dùng sử dụng độ đo tương đồng hàm ý thống kê có khả năng ứng dụng vào thực tế.

TÀI LIỆU THAM KHẢO

[1] F. Liu and H. J. Lee, Use of social network information to enhance collaborative filtering performance, Expert Systems with Applications 37(7), pp.4772-4778, 2010.

[2] F. Maxwell Harper and Joseph A. Konstan, The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19, pp.1-19, 2015.

[3] Feng Zhang, TiGong, Victor E. Lee and Gansen Zhao, Chunming Rong and Guangzhi Qu, Fast algorithms to evaluate collaborative filtering recommender systems, Knowledge-Based Systems 96 (2016) pp.96–103, 2016.

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

[5] Herlocker JL, Konstan JA, Terveen LG and Riedl JT, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems, 22(1), ISSN 1046-8188, pp.5–53, 2004.

[6] Hiep Xuan Huynh, Fabrice Guillet and Henri Briand, ARQAT: An Exploratory Analysis Tool For Interestingness Measures, pp.334-344, 2005.

[7] Jack S. Breese, David Heckerman and Carl M. Kadie, Anonymous web data from www.microsoft.com, Microsoft Research, Redmond WA, 98052-6399, USA, https://kdd.ics.uci.edu/databases/msweb/msweb.html, 1998.

[8] Martin P. Robillard, Walid Maalej, Robert J. Walker and Thomas Zimmermann, Recommendation Systems in Software Engineering, Springer Heidelberg New York Dordrecht London, ISBN 978-3-642-45135-5, 2014.

[9] Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, Collaborative Filtering Recommender Systems, Foundations and Trends in Human–Computer Interaction Vol. 4, No. 2 (2010), pp.81–173, 2010.

[10] Michael Hahsler, Lab for Developing and Testing Recommender Algorithms, Copyright (C) Michael Hahsler, http://R-Forge.R-project.org/projects/recommenderlab/, 2015.

[11] Michael Hahsler, recommenderlab: A Framework for Developing and Testing Recommendation Algorithms, the Intelligent Data Analysis Lab at SMU, http://lyle.smu.edu/IDA/recommenderlab/, 2011.

[12] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, inACM CSCW ’94, pp. 175–186, ACM, 1994.

[13] Prem Meville and Vikas Sindhwani, Recommender Systems, Encyclopedia of Machine Learning, Springer-Verlag, pp. 829-838, 2010.

[14] R. Gras and P. Kuntz, An overview of the Statistical Implicative Analysis (SIA) development, Statistical Implicative Analysis - Studies in Computational Intelligence (Volume 127), Springer-Verlag, pp.11-40, 2008.

[15] U. Shardanand and P. Maes, Social information filtering: Algorithms for automating “word of mouth”, in ACM CHI ’95, pp. 210–217, ACM Press/Addison-Wesley Publishing Co., 1995.

[16] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, Recommending and evaluating choices in a virtual community of use, inACM CHI ’95, pp. 194–201, ACM Press/Addison-Wesley Publishing Co., 1995.

USER-BASED COLLABORATIVE FILTERING RECOMMENDATION SYSTEM BASED ON IMPLICATION STATISTIC MEASURES

Phan Quoc Nghia, Nguyen Minh Ky, Dang Hoai Phuong, Huynh Xuan Hiep

ABSTRACT*— From the first appearance, recommender system in general and User-based collaborative filtering recommendation system have been developed greatly in technology and their application in life. In particular, recommender systems are used by many managers as an effective tool in order to support the business in various fields such as Amazon, Netflix and Pandora. However, the present generation of recommender systems has not fully met the requirements of users yet. In this paper, we propose a new approach for User-based collaborative filtering recommender system. The User-based collaborative filtering recommender system based on Implication statistic measures. In the system, we build a new similarity measures for two users are based on the Implication intensity measures. It is called statistical implicative similarity measures. Through experiments on two datasets MovieLense and MSWeb show that our similarity measures has fairly good results on User-based collaborative filtering model compared with traditional similarity measures as Pearson correlation, Cosine similarity, and Jaccard.*