

HỆ TƯ VẤN DỰA TRÊN ĐỘ ĐO CƯỜNG ĐỘ HÀM Ý VÀ TRÁCH NHIỆM

Phan Phương Lan¹, Huỳnh Hữu Hưng², Huỳnh Xuân Hiệp¹

¹ Trường Đại học Cần Thơ

² Trường Đại học Bách khoa Đà Nẵng, Đại học Đà Nẵng

pplan@ctu.edu.vn, hhhung@dut.udn.vn, hxhiiep@ctu.edu.vn

TÓM TẮT: Hệ tư vấn được sử dụng để dự đoán sở thích của người dùng đối với một sản phẩm hoặc dịch vụ cụ thể cũng như gợi ý các sản phẩm hoặc dịch vụ phù hợp cho người dùng. Nhiều phương pháp dùng trong khai thác dữ liệu như sự phân lớp hay luật kết hợp được áp dụng vào bài toán hệ tư vấn. Bài báo này đề xuất một mô hình tư vấn mới sử dụng các luật kết hợp và một số độ đo hàm ý thống kê quan trọng. Trong mô hình được đề xuất, các độ đo hỗ trợ và tin cậy được sử dụng để tạo ra các luật kết hợp; độ đo cường độ hàm ý kết hợp với độ đo trách nhiệm được sử dụng để lọc tập luật và xếp hạng các gợi ý. Mô hình này hoàn toàn có thể mở rộng cho những độ đo hấp dẫn khác. Hệ tư vấn dựa trên mô hình đề xuất được xây dựng và đánh giá theo bốn kịch bản và trên hai tập dữ liệu. Kết quả thực nghiệm cho thấy hệ tư vấn dựa trên độ đo cường độ hàm ý kết hợp với độ đo trách nhiệm: làm giảm kích thước của mô hình dùng để đưa ra các gợi ý nhưng vẫn duy trì được tính chính xác, qua đó làm giảm thời gian tư vấn; cho kết quả gợi ý tốt hơn các kết quả gợi ý của hệ tư vấn dựa trên cùng mô hình đề xuất nhưng khác độ đo (thống kê chi-square và hệ số Bayesian); cho kết quả gợi ý tốt hơn các kết quả gợi ý của hệ tư vấn dựa trên mô hình tư vấn AR được tích hợp trong recommenderlab sử dụng độ đo support và lift.

Từ khóa: hệ tư vấn, luật kết hợp, cường độ hàm ý, độ đo trách nhiệm.

I. GIỚI THIỆU

Sự tăng trưởng nhanh và đa dạng của các thông tin có sẵn trên internet cũng như sự phát triển của nhiều dịch vụ như thương mại điện tử, giải trí thường làm người dùng bị quá tải, dẫn đến họ có thể đưa ra các quyết định không đúng. Hệ tư vấn (recommender systems/recommendation systems - RSs) [6] là kỹ thuật hay công cụ phần mềm được nhúng trong một ứng dụng hoặc trang web mà nó sử dụng các loại dữ liệu khác nhau 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. Các hệ tư vấn được áp dụng trong nhiều lĩnh vực của cuộc sống [8] như thương mại điện tử, du lịch điện tử, học tập điện tử, dịch vụ điện tử, v.v. Một số ứng dụng nổi tiếng của RS là: hệ tư vấn của Amazon.com gợi ý cho người dùng các mặt hàng mà họ nên mua, hệ tư vấn Movie của Netflix đề xuất cho người dùng những bộ phim mà họ nên xem, v.v. Nhiều loại hệ tư vấn đã được phát triển [5][6] như hệ tư vấn: dựa trên nội dung, lọc cộng tác, dựa trên ngữ cảnh, dựa trên luật kết hợp, lai ghép, v.v. Các hệ tư vấn dựa trên nội dung sử dụng những thuộc tính mô tả (được gọi là nội dung) của các mục dữ liệu (ví dụ: bài hát, bộ phim, mặt hàng) để gợi ý các mục tương tự như các mục mà người dùng đã yêu thích trong quá khứ. Các hệ tư vấn lọc cộng tác gợi ý cho người dùng các mục dữ liệu dựa trên những mục đã được xếp hạng trước đó bởi những người dùng khác. Các hệ tư vấn dựa trên luật kết hợp được xây dựng trên tập luật và độ đo hấp dẫn để đưa ra những gợi ý cho người dùng. Các hệ tư vấn khác được phát triển dựa trên sự kết hợp của một số dạng vừa nêu và với những thông tin bổ sung như thông tin ngữ cảnh, thông tin nhóm. Việc đề xuất các mô hình tư vấn mới và cải tiến các phương pháp tư vấn hiện có luôn là một hướng nghiên cứu cốt lõi về hệ tư vấn. Nhiều phương pháp dùng trong khai phá dữ liệu được sử dụng trong bài toán tư vấn như: sự phân lớp, phân cụm, khai phá luật kết hợp, mô hình hồi quy [4][14]. So với các phương pháp khác, ưu điểm của phương pháp tư vấn dựa trên luật kết hợp là tính minh bạch [2] - nó có thể cung cấp sự giải thích sâu về các gợi ý cho người dùng. Trong các hệ tư vấn dựa trên luật kết hợp, hai độ đo được áp dụng rộng rãi để đánh giá các tập mục và tạo ra tập luật kết hợp là độ hỗ trợ (support) và độ tin cậy (confidence) [11]. Tuy nhiên, để có được những gợi ý tốt hơn, chất lượng của các luật kết hợp và cách thức xếp hạng các gợi ý nên được thực hiện thông qua một số độ đo hấp dẫn khách quan.

Phân tích hàm ý thống kê (statistical implicative analysis) [12] là một phương pháp phân tích dữ liệu được đề xuất bởi Gras. Phân tích hàm ý thống kê đã và đang được áp dụng trong nhiều lĩnh vực thực tiễn như tâm lý học, khoa học xã hội, giảng dạy, sinh tin học, v.v. Mục đích của phương pháp này là phát hiện các xu hướng trong một tập các thuộc tính (biến) bằng cách sử dụng những độ đo hàm ý thống kê. Các độ đo này được sử dụng để phát hiện những luật hoặc R-luật (luật của luật) có mối quan hệ hàm ý mạnh giữa hai vế của luật, để đo tính điển hình của một cá nhân đối với sự hình thành của một luật hoặc một R-luật, hoặc để đo tính trách nhiệm của một cá nhân đối với sự tồn tại của một luật hoặc một R-luật. Vì vậy, những độ đo hàm ý thống kê là phù hợp để phát triển các hệ tư vấn dựa trên luật kết hợp.

Bài báo này đề xuất một mô hình tư vấn mới dựa trên các luật kết hợp và một số độ đo hàm ý thống kê quan trọng nhằm dự đoán các đánh giá (/xếp hạng - rating) còn thiếu của người dùng cho các mục dữ liệu nào đó cũng như đề xuất các mục phù hợp nhất với người dùng. Trong mô hình được đề xuất, độ đo cường độ hàm ý được sử dụng để lọc tập các luật; độ đo cường độ hàm ý kết hợp với độ đo trách nhiệm được sử dụng để sắp xếp các gợi ý. Tuy nhiên, mô hình này có thể được mở rộng cho những độ đo hấp dẫn khác như hệ số Bayesian, thống kê chi-square, v.v. bằng cùng một phương pháp. Thông qua mô hình đề xuất, các hệ tư vấn sẽ được xây dựng.

Bài báo được tổ chức thành năm phần. Phần thứ nhất giới thiệu chung. Phần II mô tả ngắn gọn các độ đo hàm ý quan trọng sẽ được sử dụng, bao gồm độ đo cường độ hàm ý và độ đo trách nhiệm. Phần III đề xuất mô hình tư vấn dựa trên các độ đo đã đề cập ở trên. Phần IV trình bày các kết quả thực nghiệm và thảo luận về những kết quả đó. Phần cuối cùng là kết luận và hướng phát triển.

II. CÁC ĐỘ ĐO HÀM Ý QUAN TRỌNG

A. Độ đo cường độ hàm ý

Gọi E là một tập gồm n đối tượng hoặc cá nhân được mô tả bởi một tập hữu hạn các biến (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 (có thuộc tính b là đúng); \bar{A} (resp. \bar{B}) là phần bù của A (resp. B); $n_a = \text{card}(A)$ là số phần tử của tập A ; $n_b = \text{card}(B)$ là số phần tử của tập B ; $n_{\bar{b}} = \text{card}(\bar{B}) = n - n_b$ là số phần tử của tập \bar{B} ; và số phần tử ví dụ $n_{a\bar{b}} = \text{card}(A \cap \bar{B})$ 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 $\text{card}(X \cap \bar{Y})$ tuân theo phân phối Poisson với tham số $\lambda = \frac{n_a n_{\bar{b}}}{n}$. Kết quả, xác suất của $\text{card}(A \cap \bar{B}) \leq \text{card}(X \cap \bar{Y})$ được xác định bởi (1).

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

Với $n_{\bar{b}} \neq 0$, biến ngẫu nhiên $\text{card}(X \cap \bar{Y})$ được chuyển về biến ngẫu nhiên được chuẩn hóa $Q(a, \bar{b})$ như (2). Giá trị quan sát của $Q(a, \bar{b})$ trong thực nghiệm 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*.

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

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

Cường độ hàm ý [12] của luật $a \rightarrow b$ được định nghĩa bởi (4). Mục tiêu của độ đo này là diễn đạt chính xác tính không chắc chắn (unlikelihood) của $n_{a\bar{b}}$ trong E .

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

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

B. Độ đo trách nhiệm

Đề đo sự đóng góp của một cá nhân đối với sự tồn tại của một luật hay một R-luật, *độ đo trách nhiệm contribution* đã được phát triển [12]. Trong bài báo này, chúng tôi tập trung vào việc đo sự góp phần của một đối tượng i đối với sự hình thành luật có dạng $a \rightarrow b$. Gọi $a(i)$ (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 , độ đo trách nhiệm được xác định bởi (5).

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

Trong thực tế, p được đặt bằng giá trị trung lập 0,5.

III. MÔ HÌNH TƯ VẤN DỰA TRÊN CÁC ĐỘ ĐO CƯỜNG ĐỘ HÀM Ý VÀ TRÁCH NHIỆM

Mô hình tư vấn (hay mô hình hệ tư vấn) được đề xuất sẽ sử dụng ma trận đánh giá (/ma trận xếp hạng - rating matrix) làm dữ liệu đầu vào chính, sau đó trả về các đánh giá được dự đoán và danh sách gồm N mục có thứ hạng cao nhất được gợi ý cho người dùng. Phương pháp tư vấn này sử dụng mô hình phụ thuộc giữa các mục được mô tả bởi một tập các luật kết hợp và một số độ đo hàm ý quan trọng để đưa ra gợi ý. Tập các luật kết hợp được xây dựng bằng cách khai phá ma trận đánh giá thông qua sự kết hợp của hai độ đo hỗ trợ và tin cậy. Điểm mạnh của cách tiếp cận dựa trên luật kết hợp so với các phương pháp khác là tính minh bạch - nghĩa là có thể giải thích sâu về các gợi ý cho người dùng.

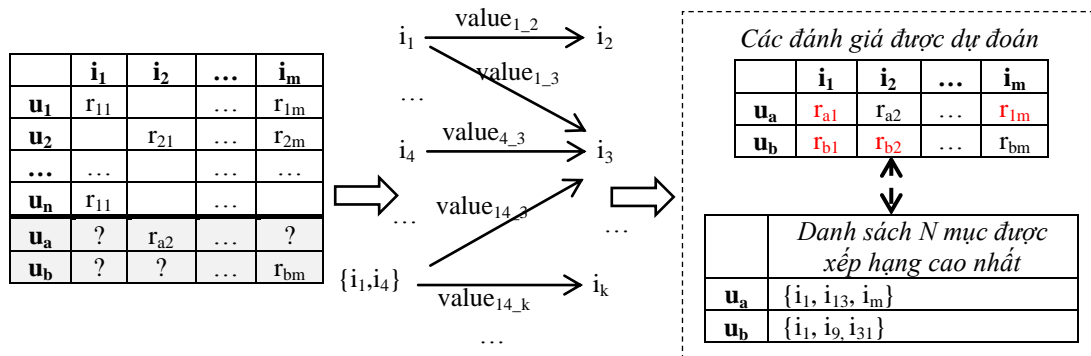
Như bài toán tư vấn hai chiều [7], mô hình tư vấn được đề xuất có thể được hình thức hóa như sau:

- 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 (item). Mục có thể là bài hát, bộ phim, mặt hàng, v.v.
- Gọi $R = (r_{jk})$ trong đó $j = 1..n$ và $k = 1..m$ là một ma trận đánh giá lưu trữ thông tin phản hồi của người dùng về các mục. r_{jk} ghi lại đánh giá/xếp hạng (rating) cho một mục i_k được thực hiện bởi một người dùng u_j .
- Gọi $f: U \times I \rightarrow R$ là một hàm ánh xạ từ những kết hợp của người dùng và các mục vào các đánh giá r .

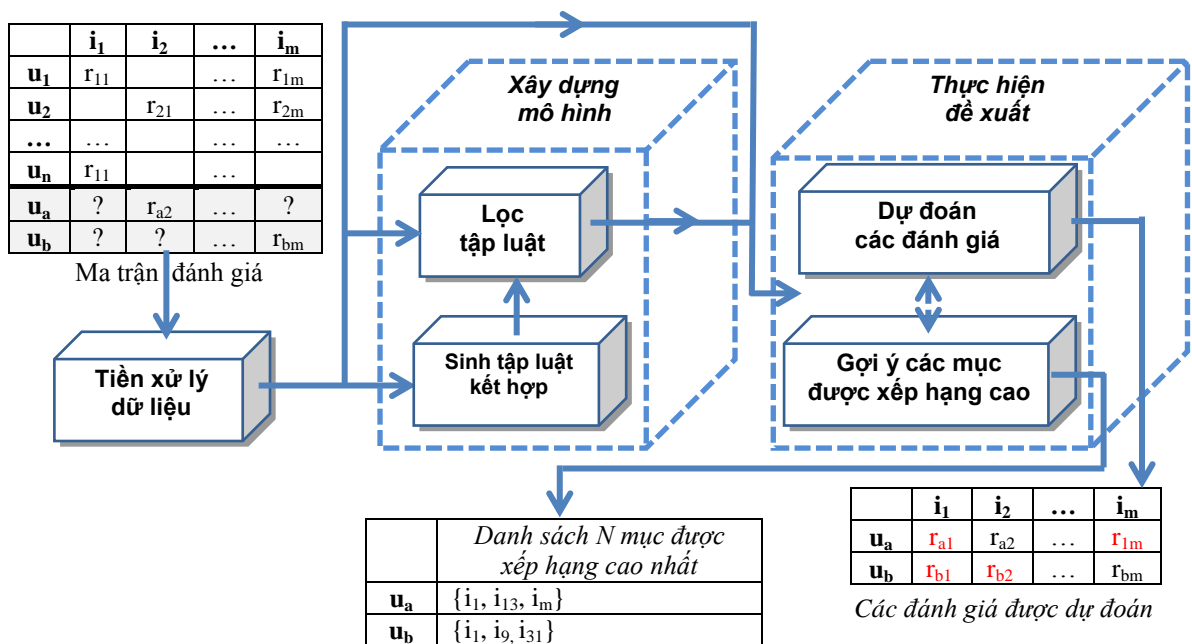
Mục tiêu của mô hình đề xuất là tìm một hàm $f': U \times I \rightarrow R'$ sao cho hàm $\xi(r, r')$ đạt được kết quả tốt hơn (qua các độ đo như: chính xác precision và bao phủ recall) cũng như làm giảm thời gian dự đoán để đưa ra các gợi ý khi so sánh với một số phương pháp khác. Mô hình đề xuất sử dụng các độ đo hỗ trợ, tin cậy, cường độ hàm ý và độ đo trách nhiệm để tìm ra hàm f' .

Thành phần cốt lõi của các phương pháp tư vấn là những độ đo được sử dụng để: tìm những người dùng tương tự, tìm các sản phẩm tương tự, khám phá các mối quan hệ với sự gắn kết cao, hoặc xếp hạng các gợi ý, v.v. Trong mô hình đề xuất, các độ đo tin cậy và hỗ trợ được sử dụng để tạo ra một tập các luật; độ đo cường độ hàm ý được sử dụng để lọc tập các luật; và cường độ hàm ý kết hợp với độ đo trách nhiệm được sử dụng để xếp hạng các gợi ý. Dữ liệu đầu vào chính của mô hình được đề xuất là một ma trận đánh giá lưu trữ thông tin quan trọng được tạo ra trong quá trình tương tác giữa người dùng và hệ thống, chẳng hạn như đánh giá/xếp hạng của người dùng cho các bộ phim đã xem, hay các bài hát đã nghe. Các đánh giá được thu thập một cách tường minh hoặc không tường minh; có thể được định nghĩa theo nhiều cách: đánh giá liên tục, đánh giá trên khoảng, đánh giá thứ tự, đánh giá nhị phân hoặc đánh giá nhất phân. Ma trận đánh giá có thể có các ô r_{jk} trống (những đánh giá bị thiếu) nếu người dùng u_j chưa đánh giá mục i_k hoặc không biết về nó. Mô hình tư vấn đề xuất được phác họa như Hình 1.



$value_{i,j}$ là giá trị của luật $i \rightarrow j$ theo một độ đo xác định như: độ tin cậy, độ hỗ trợ, cường độ hàm ý.

Hình 1. Bản phác thảo của mô hình tư vấn dựa trên các độ đo cường độ hàm ý và trách nhiệm



Hình 2. Các bước xử lý chính của mô hình tư vấn dựa trên các độ đo cường độ hàm ý và trách nhiệm

Quá trình đưa ra các gợi ý của mô hình đề xuất được trình bày trong Hình 2. Trong đó, bước 2 và bước 3 được sử dụng để học mô hình gợi ý – mô hình được mô tả bằng một tập các luật kết hợp. Bước 4 và bước 5 sử dụng mô hình đã học để đưa ra các gợi ý. Hai bước cuối cùng (bước 4 và 5) có nhiều công việc con tương tự nhau, hơn nữa giá trị của các đánh giá được dự đoán ở dạng nhị phân (1 hoặc 0) nên kết xuất “Danh sách N mục được xếp hạng cao nhất” được đặc biệt quan tâm. Quá trình xử lý bao gồm năm bước:

- Tiền xử lý dữ liệu. Bước này bao gồm hai công việc con: (1) lựa chọn dữ liệu có liên quan (trong tập dữ liệu) bằng cách xác định số lượng tối thiểu các mục được đánh giá bởi mỗi người dùng và số lượng tối thiểu những người dùng thực hiện đánh giá trên mỗi mục; (2) nhị phân hóa tập dữ liệu đã được chọn lại nếu nó là không phải là dữ liệu nhị phân. Kết quả của bước này là tập dữ liệu *dataset*.
- Sinh tập các luật kết hợp *ruleset* từ tập dữ liệu *dataset*. Bước này bao gồm hai công việc con: (1) lựa chọn ngưỡng cho hai độ đo tin cậy và hỗ trợ; (2) tạo ra một tập các luật kết hợp dựa trên các ngưỡng được lựa chọn bằng giải thuật apriori. Tập luật này bao gồm tất cả những luật có độ dài tối đa là l và tối thiểu là 2. Nếu luật có độ dài l thì nó sẽ có $(l - 1)$ mục bên về trái và 1 mục bên về phải.
- Lọc tập luật *ruleset* dựa trên độ đo cường độ hàm ý. Bước này bao gồm ba công việc con: (1) biểu diễn mỗi luật bằng một bộ bản số gồm bốn giá trị $\{n, n_a, n_b, n_{ab}\}$; (2) tính giá trị cường độ hàm ý của từng luật; (3) lựa chọn ngưỡng thích hợp của cường độ hàm ý nhằm giảm kích thước của tập luật. Thuật toán được sử dụng trong các công việc con 1 và 2 đã được chúng tôi trình bày trong [9].
- Đề xuất N mục phù hợp nhất cho những người dùng trong tập dữ liệu *recset*. Bước này bao gồm hai công việc con: (1) với mỗi người dùng, tìm tất cả các luật có về trái là tập con của tập các mục đã được đánh giá bởi người dùng đó; (2) xếp hạng (sắp xếp) các luật được lọc theo giá trị cường độ hàm ý hoặc sự kết hợp giữa cường độ hàm ý với độ đo trách nhiệm, và trình bày N về phải duy nhất của các luật đã được sắp xếp cho người dùng. Thuật toán thô được sử dụng tại bước này như sau.

Input: tập luật *ruleset* có kích thước r luật, tập dữ liệu *recset* của p người dùng cần được gợi ý với nhiều đánh giá (của người dùng cho các mục dữ liệu) bị thiếu.

Output: danh sách các mục được đề xuất cho những người dùng cần được gợi ý

Các bước thực hiện:

- Tạo ma trận *ruleleft* có kích thước (m mục x r luật) từ tập *ruleset*. Trong đó, $ruleleft[i, j]=1/0$ nếu mục thứ i trong tập mục I có xuất hiện/không xuất hiện ở về trái của luật thứ j .
- Tạo ma trận *ruleuserleft* có kích thước (r luật x p người dùng) từ tập *ruleleft* và *recset*. Trong đó, $ruleuserleft[i, j]=1/0$ nếu về trái của luật thứ i gồm những mục được thích/không được thích (hay không được biết đến) bởi người dùng thứ j .
- For từng người sử dụng i trong tập *recset* do
 - Lọc tất cả các luật j mà $ruleuserleft[j, i]=1$.
 - Sắp xếp tất cả các luật đó theo giá trị độ đo hấp dẫn giảm dần.
 - Lọc tất cả các về phải (các mục) không trùng nhau của các luật đã sắp xếp.
 - Đề xuất N mục được xếp cao nhất.

- Dự đoán các xếp hạng/đánh giá bị thiếu. Bước này bao gồm hai công việc con: (1) với mỗi đánh giá bị thiếu ở hàng i và cột j , tìm tất cả các luật có về trái là tập con của tập các mục đã được đánh giá bởi người dùng i và về phải là mục j ; (2) xếp hạng (sắp xếp) các luật được lọc theo giá trị cường độ hàm ý hoặc sự kết hợp giữa cường độ hàm ý và độ đo trách nhiệm, sau đó gán giá trị 0 hoặc 1 cho đánh giá bị thiếu dựa trên một ngưỡng xác định. Thuật toán được sử dụng tại bước này có nhiều hoạt động tương tự như bước trên.

Mô hình được đề xuất trong bài báo này hoàn toàn có thể sử dụng những độ đo hấp dẫn khác như hệ số Bayesian, thống kê chi-square vì mỗi luật được biểu diễn bằng một bộ bản số gồm bốn giá trị $\{n, n_a, n_b, n_{ab}\}$. Do đó, giá trị của các luật theo một độ đo cụ thể sẽ được xác định và được dùng để lọc các luật cũng như sắp xếp các gợi ý.

IV. THỰC NGHIỆM

A. Thiết lập thực nghiệm

1. Dữ liệu thực nghiệm

Hai tập dữ liệu được sử dụng trong thực nghiệm là MSWeb [1] và MovieLens¹. 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 liệt kê tất cả các vùng của trang web (Vroots) được truy cập bởi người dùng đó. MSWeb bao gồm: 32710 người dùng, 285 Vroot và 98653 đánh giá với giá trị TRUE/1. Tập dữ liệu MovieLens - lưu đánh giá của người dùng cho các bộ phim - đã được thu thập thông qua trang web movielens.umn.edu trong khoảng thời gian bảy tháng. Tập dữ liệu chứa: 943 người dùng, 1664

¹ <http://www.grouplens.org>

phim và 99392 đánh giá với các giá trị từ 1 đến 5. Để có thể sử dụng mô hình đề xuất, tập dữ liệu MovieLens phải được nhị phân hóa. Vì mỗi đánh giá nhận một trong các giá trị từ 1* đến 5* nên ngưỡng để nhị phân hóa là 3; khi đó, các đánh giá được gán giá trị 1 nếu chúng lớn hơn hoặc bằng ngưỡng 3 và 0 nếu ngược lại.

2. Công cụ thực nghiệm

Công cụ Interestingnesslab (được chúng tôi trình bày trong [10]) được cải tiến để phát triển và đánh giá mô hình tư vấn được đề xuất. Các thực nghiệm được chạy trên máy tính có cấu hình hệ điều hành Window 8, RAM 16 GB, và bộ vi xử lý Intel Pentium G630 2.7GHz. Bên cạnh đó, chúng tôi còn sử dụng mô hình tư vấn dựa trên luật (AR) của gói recommenderlab [15] để so sánh với mô hình đề xuất.

3. Các kịch bản thực nghiệm

Bốn kịch bản thực nghiệm được đề xuất là:

- Xác định các ngưỡng thích hợp cho độ tin cậy, độ hỗ trợ và cường độ hàm ý nhằm giảm kích thước của mô hình được dùng để đưa ra các gợi ý nhưng vẫn duy trì được tính chính xác. Kịch bản này khám phá các ngưỡng của độ tin cậy và hỗ trợ dùng để tạo ra tập luật kết hợp và của cường độ hàm ý dùng để lọc tập luật. Mục đích của kịch bản này là vẫn duy trì tính chính xác của mô hình tư vấn nhưng giảm kích thước của tập luật (số lượng luật), do đó làm giảm thời gian tư vấn.
- Đánh giá kết quả gợi ý của hệ tư vấn dựa trên cường độ hàm ý so với các kết quả gợi ý của hai hệ tư vấn dựa trên hệ số Bayesian và thống kê chi-square tương ứng. Những hệ thống này sử dụng cùng một mô hình đề xuất nhưng các độ đo khác nhau để lọc và xếp hạng các gợi ý.
- Đánh giá kết quả gợi ý của hệ tư vấn dựa trên cường độ hàm ý kết hợp với độ đo trách nhiệm so với các kết quả gợi ý của hai hệ tư vấn dựa trên cường độ hàm ý và hệ số Bayesian tương ứng. Tương tự như kịch bản trên, ba hệ thống này sử dụng cùng một mô hình đề xuất nhưng các độ đo khác nhau để lọc và xếp hạng các gợi ý.
- Đánh giá kết quả gợi ý của hệ tư vấn dựa trên cường độ hàm ý kết hợp với độ đo trách nhiệm so với kết quả gợi ý của hai hệ tư vấn dựa trên độ đo hỗ trợ support và độ đo lift tương ứng. Hai hệ thống sau sử dụng mô hình dựa trên luật kết hợp được tích hợp trong gói recommenderlab.

4. Đánh giá hệ tư vấn

Để đánh giá các hệ tư vấn, tập dữ liệu đầu vào được phân tách thành tập dữ liệu huấn luyện có kích thước (n người dùng \times m mục dữ liệu) và tập dữ liệu kiểm thử có kích thước (p người dùng \times m mục dữ liệu). Tập dữ liệu huấn luyện được sử dụng để xây dựng mô hình trong khi tập dữ liệu kiểm thử được sử dụng để kiểm tra mô hình. Tập dữ liệu kiểm thử lại được chia thành hai phần: tập dữ liệu truy vấn và tập dữ liệu đích có cùng kích thước (p người dùng \times m mục dữ liệu). Tuy nhiên, với mỗi người dùng, tập dữ liệu truy vấn chỉ có *given* đánh giá (mục dữ liệu được đánh giá) được chọn ngẫu nhiên; tập dữ liệu đích gồm những đánh giá còn lại. Tập dữ liệu truy vấn được dùng để đưa ra các gợi ý dựa trên mô hình đã học còn tập dữ liệu đích được dùng để đánh giá (tính chính xác, tính bao phủ) của 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 để học mô hình và 1 tập con còn lại được sử dụng để kiểm tra. Trong thực nghiệm này, k được chọn là 4. Khi số lượng các gợi ý (độ dài) 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 [3]. Ngoài ra, đường cong ROC được sử dụng để so sánh tính hiệu quả 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 giải thuật đó là hiệu quả hơn. Đường cong Precision - Recall được xây dựng dựa trên độ chính xác (precision) và độ bao phủ (recall). Đường cong ROC được xây dựng dựa trên độ nhạy (còn được gọi là True Positive Rate - TPR) và phần bù của độ đặc hiệu specificity (còn được gọi là False Positive Rate - FPR). Những phép đo này được xây dựng dựa vào ma trận nhầm lẫn như Bảng 1 và có công thức tính như (6), (7) và (8).

Bảng 1. Ma trận nhầm lẫn

Thực tế / Gợi ý	Không được gợi ý	Được gợi ý
Không được thích	TN	FP
Được thích	FN	TP

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall/TPR = \frac{TP}{TP + FN} \quad (7)$$

$$FPR = 1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} \quad (8)$$

5. Tiền xử lý dữ liệu

Để tăng tính chính xác trong việc xây dựng mô hình gợi ý và kiểm chứng nó, 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 dùng chỉ đánh giá một vài mục thì các đánh giá có thể bị thiên vị. Do đó, để chọn dữ liệu có liên quan trong tập dữ liệu, các phân vị cho một tập các số mục dữ liệu trên mỗi người dùng và cho một tập các số người dùng trên mỗi mục dữ liệu phải được xem xét. Phân vị p có giá trị x_p sao cho $p\%$ của tất cả các giá trị của x nhỏ hơn x_p . Sau khi phân tích phân vị, với tập dữ liệu MSWeb, số lượng người dùng đã xem từ trên 10 mục website (số Vroot) và số lượng mục website được xem bởi 50 người dùng trở lên được chọn để trích xuất dữ liệu có liên quan; với tập dữ liệu MovieLens sau khi được nhị phân hóa, số lượng người dùng đã đánh giá từ trên 50 bộ phim và số lượng phim được đánh giá bởi 100 người dùng trở lên được chọn để trích xuất dữ liệu.

Thông tin chung về hai tập MSWeb và MovieLens sau khi lọc dữ liệu được trình bày trong Bảng 2, Bảng 3 và Bảng 4. Bảng 2 và Bảng 3 cho thấy 0% các giá trị của số mục website (Vroot) trên mỗi người dùng nhỏ hơn 7 và 0% các giá trị của số phim trên mỗi người dùng nhỏ hơn 14. Như vậy, số lượng Vroot và số lượng phim tối đa (số *given*) trên mỗi người dùng trong tập dữ liệu kiểm thử được sử dụng để xây dựng các gợi ý là 7 và 14 tương ứng.

Bảng 2. Các phân vị của tập dữ liệu MSWeb sau khi được tiền xử lý

	0%	10%	40%	60%	80%	90%	100%
Số mục của website (Vroot) trên mỗi người dùng	7	10	11	12	14	15.6	34
Số người dùng trên mỗi mục của website (Vroot)	2	8.4	24	44.4	97.8	191	604

Bảng 3. Các phân vị của tập dữ liệu MovieLens sau khi được tiền xử lý

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Số phim trên mỗi người dùng	14	37	43	51	63	77	92	104.3	125	153	248
Số người dùng trên mỗi phim	37	78	90.2	100	112	126	145	170	198	245	435

Bảng 4. Thông tin chung của hai tập dữ liệu MSWeb và MovieLens sau khi được tiền xử lý

Tập dữ liệu	Số dòng (người dùng)	Số cột (vùng của website/ phim)	Số đánh giá	Số $maxgiven^2$
MSWeb	875	135	10487	7
MovieLens	560	332	48266	14

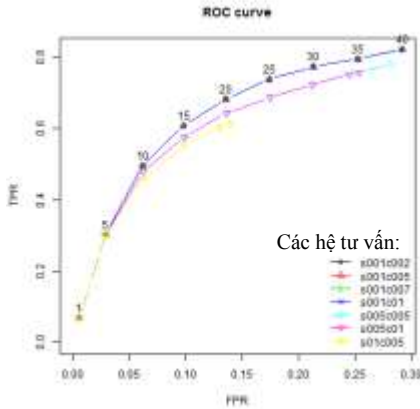
B. Kịch bản 1: chọn các ngưỡng phù hợp để giảm kích thước của mô hình dùng để đưa ra các gợi ý nhưng vẫn đảm bảo tính chính xác

1. Kết quả thực nghiệm

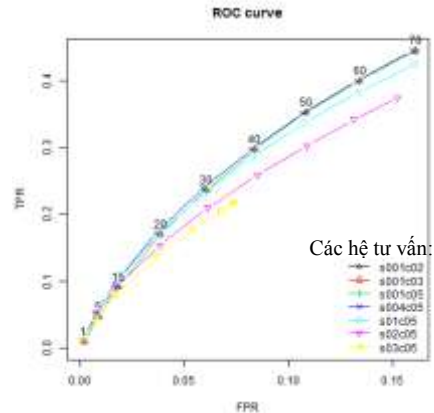
Để chọn các ngưỡng thích hợp cho độ tin cậy và hỗ trợ, đường cong ROC và đường cong precision-recall được sử dụng khi so sánh các kết quả gợi ý. Bên cạnh đó, để không bỏ sót các luật kết hợp với chất lượng cao, ngưỡng hỗ trợ và ngưỡng tin cậy nên được gán các giá trị nhỏ. Sử dụng mô hình đề xuất, các hệ tư vấn có các ngưỡng hỗ trợ và tin cậy khác nhau được xây dựng. Hình 3 hiển thị đường cong ROC của các hệ tư vấn (s001c002, s001c005, s001c007, s001c01, s005c005, s005c01, s01c005) sử dụng: tập dữ liệu MSWeb với các ngưỡng (hỗ trợ, tin cậy) lần lượt là (0.01, 0.02), (0.01, 0.05), (0.01, 0.07), (0.01, 0.1), (0.05, 0.05), (0.05, 0.1) và (0.1, 0.05); số lượng Vroot trên mỗi người dùng trong tập kiểm thử được sử dụng để xây dựng các gợi ý (số *given*) là 7; và độ dài luật tối đa là 4. Khi số *given* và độ dài luật được thay đổi, các hệ tư vấn cũng cho kết quả tương tự như Hình 3. Hình 4 hiển thị đường cong ROC của các hệ tư vấn (s001c02, s001c03, s001c05, s004c05, s01c05, s02c05, s03c05) sử dụng: tập dữ liệu MovieLens với các ngưỡng (hỗ trợ, tin cậy) lần lượt là (0.01, 0.2), (0.01, 0.3), (0.01, 0.5), (0.04, 0.5), (0.1, 0.5), (0.2, 0.5) và (0.3, 0.5); số lượng phim trên mỗi người dùng trong tập kiểm thử được sử dụng để xây dựng các gợi ý (số *given*) là 14; và độ dài luật tối đa là 3.

Để cải thiện thời gian tư vấn, số lượng luật trong tập luật nên được thu giảm dựa trên một ngưỡng thích hợp cho độ đo cường độ hàm ý. Trong kịch bản này, mô hình đề xuất được sử dụng để xây dựng các hệ tư vấn có cùng ngưỡng hỗ trợ và tin cậy nhưng có ngưỡng cường độ hàm ý khác nhau. Bảng 5 trình bày kết quả của hai hệ tư vấn cùng sử dụng tập dữ liệu MSWeb, cùng số Vroot để xây dựng các gợi ý 5, cùng ngưỡng hỗ trợ 0.01 và ngưỡng tin cậy 0.1 nhưng một hệ thống lấy tất cả các luật (ngưỡng cường độ hàm ý=0.0) và một hệ thống sử dụng các luật có cường độ hàm ý từ 0.90 trở lên (ngưỡng=0.90). Ngoài ra, thực nghiệm còn cho thấy, đường cong ROC và đường cong precision-recall của hai hệ thống này gần như trùng khớp nhau hoàn toàn.

² Số lượng mục tối đa trên mỗi người dùng trong tập kiểm thử được sử dụng để xây dựng các gợi ý (theo mô hình đã học).



Hình 3. Đường cong ROC của các hệ tư vấn sử dụng tập dữ liệu MSWeb



Hình 4. Đường cong ROC của các hệ tư vấn sử dụng tập dữ liệu MovieLens

Với cách làm tương tự cho tập dữ liệu MovieLens, đường cong ROC và đường cong precision-recall của hai hệ tư vấn sử dụng ngưỡng cường độ hàm ý 0.0 và 0.9 là gần như trùng khớp nhau hoàn toàn.

Bảng 5. Số luật được sinh ra và thời gian dự đoán của hai hệ tư vấn sử dụng ngưỡng cường độ hàm ý khác nhau trên tập MSWeb.

Độ dài tối đa của luật	Số luật		Thời gian dự đoán (sec)	
	Ngưỡng cường độ hàm ý=0.0	Ngưỡng cường độ hàm ý=0.9	Ngưỡng cường độ hàm ý=0.0	Ngưỡng cường độ hàm ý=0.9
2	2004	581	4.5125	4.4675
5	83057	26033	6.3900	5.2650
7	109312	37404	7.2275	5.7250

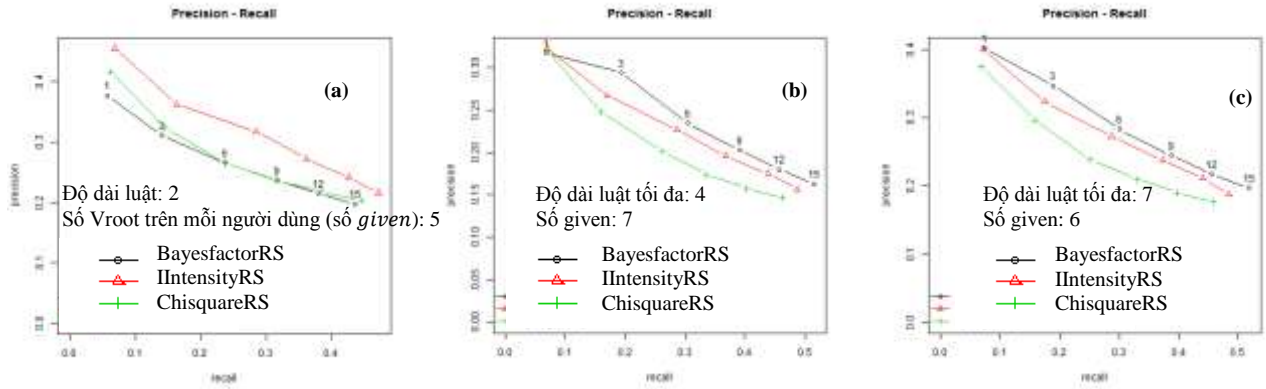
2. Thảo luận

Đối với tập dữ liệu MSWeb, Hình 3 cho thấy diện tích dưới các đường cong ROC của các hệ tư vấn s001002, s001005, s001007, và s001c01 đều bằng nhau nhưng lớn hơn các diện tích nằm dưới các đường cong ROC khác. Bên cạnh đó, với cùng ngưỡng hỗ trợ 0.01, kích thước của tập luật được tạo ra bởi ngưỡng tin cậy 0.02 hoặc 0.05 hoặc 0.07 lớn hơn kích thước của tập luật được tạo ra bởi ngưỡng tin cậy 0.1, nên trong trường hợp này, các ngưỡng hỗ trợ và tin cậy được chọn là 0.01 và 0.1 tương ứng. Ngoài ra, Bảng 5 cho thấy số luật và thời gian dự đoán của hệ thống sử dụng ngưỡng cường độ hàm ý 0.9 là nhỏ hơn so với hệ thống không sử dụng ngưỡng trong khi các đường cong ROC và đường cong precision-recall của cả hai hệ thống gần như hoàn toàn trùng khớp nhau. Vì vậy, ngưỡng phù hợp nhất của độ đo cường độ hàm ý là 0.9. Đối với tập dữ liệu MovieLens, Hình 4 cho thấy hệ tư vấn sử dụng ngưỡng hỗ trợ 0.04 và ngưỡng tin cậy 0.5 có độ chính xác cao hơn các hệ thống khác. Tuy nhiên, với các ngưỡng đã chọn này, số lượng luật là quá lớn nếu độ dài luật tối đa là lớn. Ví dụ, nếu độ dài luật là 4 và ngưỡng tin cậy là 0.04, số luật được sinh ra là: 112727547 khi sử dụng ngưỡng hỗ trợ 0.04, 5891352 khi sử dụng ngưỡng hỗ trợ 0.1, 99671 khi sử dụng ngưỡng hỗ trợ 0.2. Máy tính được sử dụng trong các thực nghiệm này sẽ bị quá tải khi số luật quá lớn. Do đó, đối với tập dữ liệu MovieLens, các ngưỡng được chọn là: 0.2 cho độ hỗ trợ và 0.5 cho độ tin cậy nếu độ dài tối đa của luật lớn hơn 2; và 0.04 cho độ hỗ trợ và 0.5 cho độ tin cậy nếu độ dài luật bằng 2.

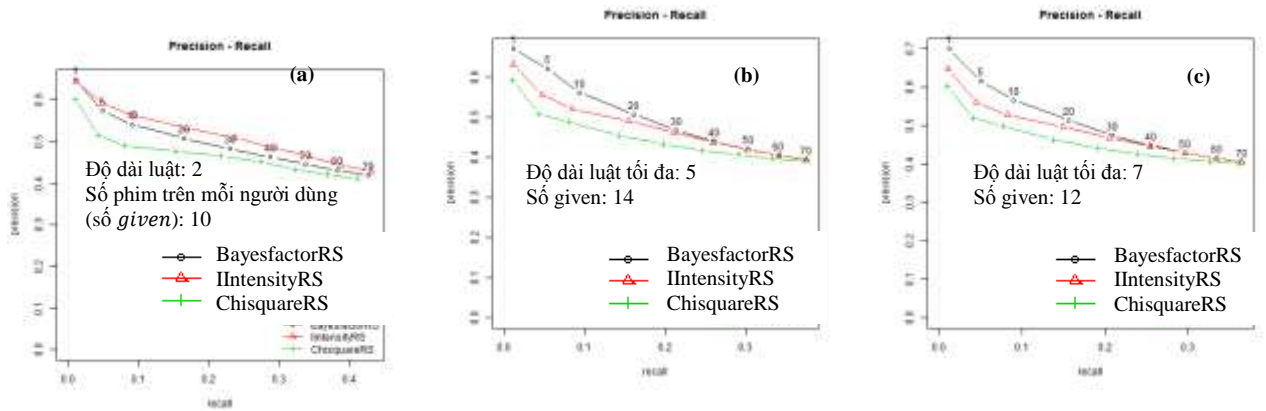
C. Kịch bản 2: đánh giá kết quả gợi ý của hệ tư vấn dựa trên cường độ hàm ý

1. Kết quả thực nghiệm

Tính chính xác và tính bao phủ của các kết quả gợi ý của ba hệ tư vấn (IntensityRS, BayesfactorRS, ChisquareRS) sử dụng độ đo cường độ hàm ý, hệ số bayesian và thống kê chi-square được thể hiện trong Hình 5 nếu dùng tập dữ liệu MSWeb, Hình 6 nếu dùng tập dữ liệu MovieLens. Hình 5 hiển thị các đường cong precision-recall của ba hệ tư vấn trong các trường hợp (độ dài luật tối đa, số Vroot trên mỗi người dùng trong tập kiểm thử được sử dụng để đưa ra gợi ý - given) lần lượt là (2,5), (4,7) và (7,6) và số Vroot cần gợi ý cho mỗi người dùng là 1, 3, 6, 9, 12, và 15. Tương tự như vậy, Hình 6 hiển thị các đường cong precision-recall của ba hệ tư vấn trong các trường hợp (độ dài luật tối đa, số phim trên mỗi người dùng trong tập kiểm thử được sử dụng để đưa ra gợi ý - given) lần lượt là (2,10), (5,14) và (7,12) và số phim cần gợi ý cho mỗi người dùng là 1, 5, 10, 20, 30, 40, 50, 60 và 70. Với độ dài luật bằng hai, khi số phim hay Vroot trên mỗi người dùng được thay đổi, các đường cong precision-recall của ba hệ tư vấn IntensityRS, BayesfactorRS, ChisquareRS là tương tự như Hình 5.a (cho tập dữ liệu MSWeb) và Hình 6.a (cho tập dữ liệu MovieLens). Với độ dài luật tối đa lớn hơn hai, khi số given được thay đổi, các đường cong precision-recall của ba hệ tư vấn IntensityRS, BayesfactorRS, ChisquareRS là cũng tương tự như Hình 5.b, 5.c (cho tập dữ liệu MSWeb) và Hình 6.b, 6.c (cho tập dữ liệu MovieLens).



Hình 5. Đường cong precision-recall của các hệ tư vấn BayesfactorRS, IIntensityRS, ChisquareRS sử dụng tập dữ liệu MSWeb



Hình 6. Đường cong precision-recall của ba hệ tư vấn BayesfactorRS, IIntensityRS, ChisquareRS sử dụng tập dữ liệu MovieLens

2. Thảo luận

Bảng 2 cho thấy 10% số người dùng đã xem ít hơn 10 Vroot và khoảng 90% số người dùng đã xem ít hơn 16 Vroot. Nếu số Vroot trên mỗi người dùng trong tập kiểm thử được sử dụng để đưa ra gợi ý (số given) là 5, 6, 7 và xét ở mức 90% như vừa nêu thì đoạn cần quan tâm trên các đường cong precision-recall là 5-11, 4-10 và 3-9 tương ứng. Tương tự như vậy cho tập dữ liệu MovieLens. Kết quả thực nghiệm cho thấy: (1) với độ dài luật bằng hai, kết quả gợi ý của hệ tư vấn sử dụng độ đo cường độ hàm ý là vượt trội hơn hai hệ tư vấn sử dụng hệ số bayesian và thống kê chi-square, cụ thể: tính chính xác và tính bao phủ là cao hơn; (2) với độ dài luật tối đa lớn hơn hai, kết quả gợi ý của hệ tư vấn sử dụng độ đo cường độ hàm ý là vượt trội hơn của hệ tư vấn sử dụng thống kê chi-square nhưng kém hơn so với của hệ thống sử dụng hệ số bayesian trên một số đoạn cần quan tâm, tuy nhiên sự khác biệt càng nhỏ dần khi số mục cần gợi ý tăng.

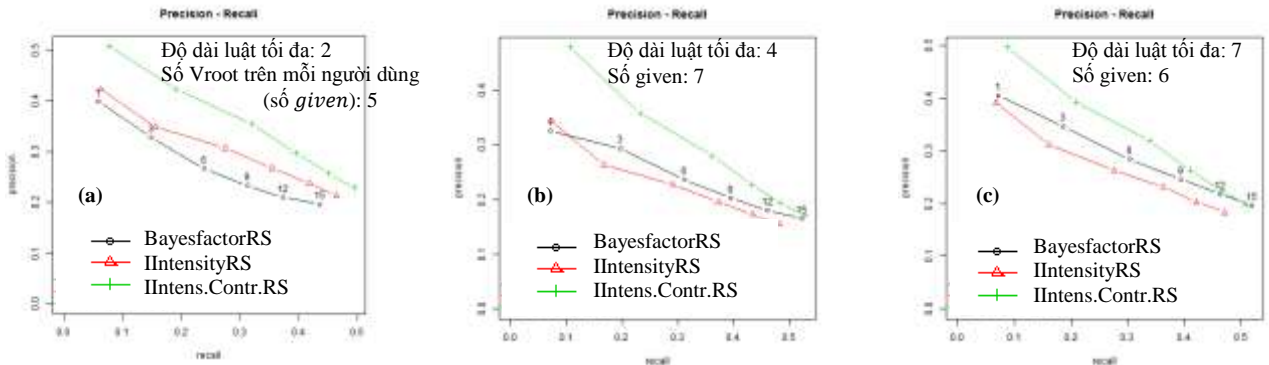
D. Kịch bản 3: đánh giá kết quả gợi ý của hệ tư vấn dựa trên sự kết hợp của cường độ hàm ý và độ đo trách nhiệm

1. Kết quả thực nghiệm

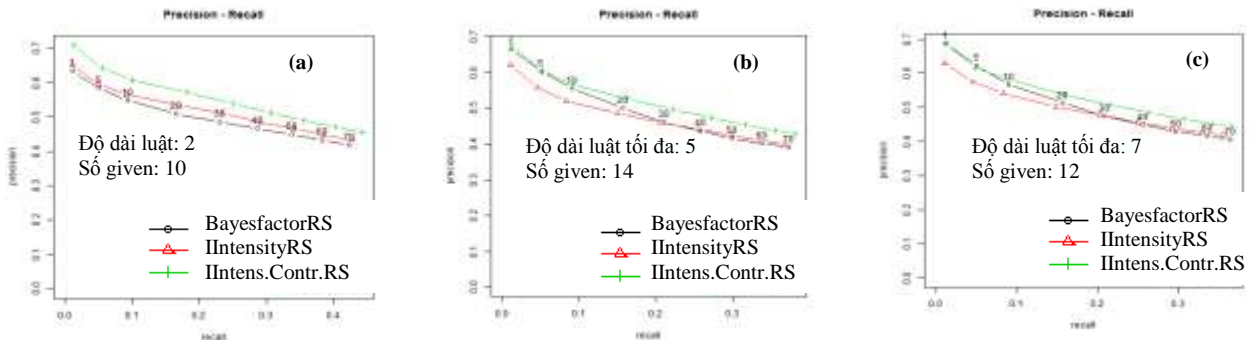
Tính chính xác và tính bao phủ của các kết quả gợi ý của ba hệ tư vấn (IIntens.Contr.RS, IIntensityRS, BayesfactorRS) sử dụng độ đo cường độ hàm ý kết hợp với độ đo trách nhiệm, cường độ hàm ý, và hệ số bayesian được thể hiện trong Hình 7 nếu dùng tập dữ liệu MSWeb, Hình 8 nếu dùng tập dữ liệu MovieLens. Hình 7 hiển thị các đường cong precision-recall của ba hệ tư vấn trong các trường hợp (độ dài luật tối đa, số vùng của website (số Vroot) trên mỗi người dùng trong tập kiểm thử được sử dụng để đưa ra các gợi ý - given) lần lượt là (2,5), (4,7) và (7,6) và số Vroot cần gợi ý cho mỗi người dùng là 1, 3, 6, 9, 12, và 15. Tương tự như vậy, Hình 8 hiển thị các đường cong precision-recall của ba hệ tư vấn trong các trường hợp (độ dài luật tối đa, số phim trên mỗi người dùng trong tập kiểm thử được sử dụng để đưa ra gợi ý - given) lần lượt là (2,10), (5,14) và (7,12) và số phim cần gợi ý cho mỗi người dùng là 1, 5, 10, 20, 30, 40, 50, 60 và 70. Khi độ dài luật tối đa và số given thay đổi, các đường cong precision-recall của ba hệ tư vấn IIntens.Contri.RS, IIntensityRS, BayesfactorRS trên tập dữ liệu MSWeb có dạng tương tự như Hình 7 và trên tập dữ liệu MovieLens có dạng tương tự như Hình 8.

2. Thảo luận

Kết quả thực nghiệm cho thấy với bất cứ độ dài luật tối đa nào, kết quả gợi ý của hệ tư vấn dựa trên cường độ sẽ được cải thiện (tốt hơn) nếu được kết hợp với độ đo trách nhiệm so với kết quả của hệ tư vấn dựa trên hệ số bayesian. Cụ thể, tính chính xác (tỷ lệ số mục dữ liệu vừa được gợi ý vừa được yêu thích bởi người dùng trên tổng số mục được gợi ý) cao hơn cũng như tính bao phủ (tỷ lệ số mục dữ liệu vừa được gợi ý vừa được yêu thích bởi người dùng trên tổng số mục được yêu thích) cao hơn.



Hình 7. Đường cong precision-recall của các hệ tư vấn BayesfactorRS, IIntensityRS, IIntens.Contr.RS sử dụng tập dữ liệu MSWeb

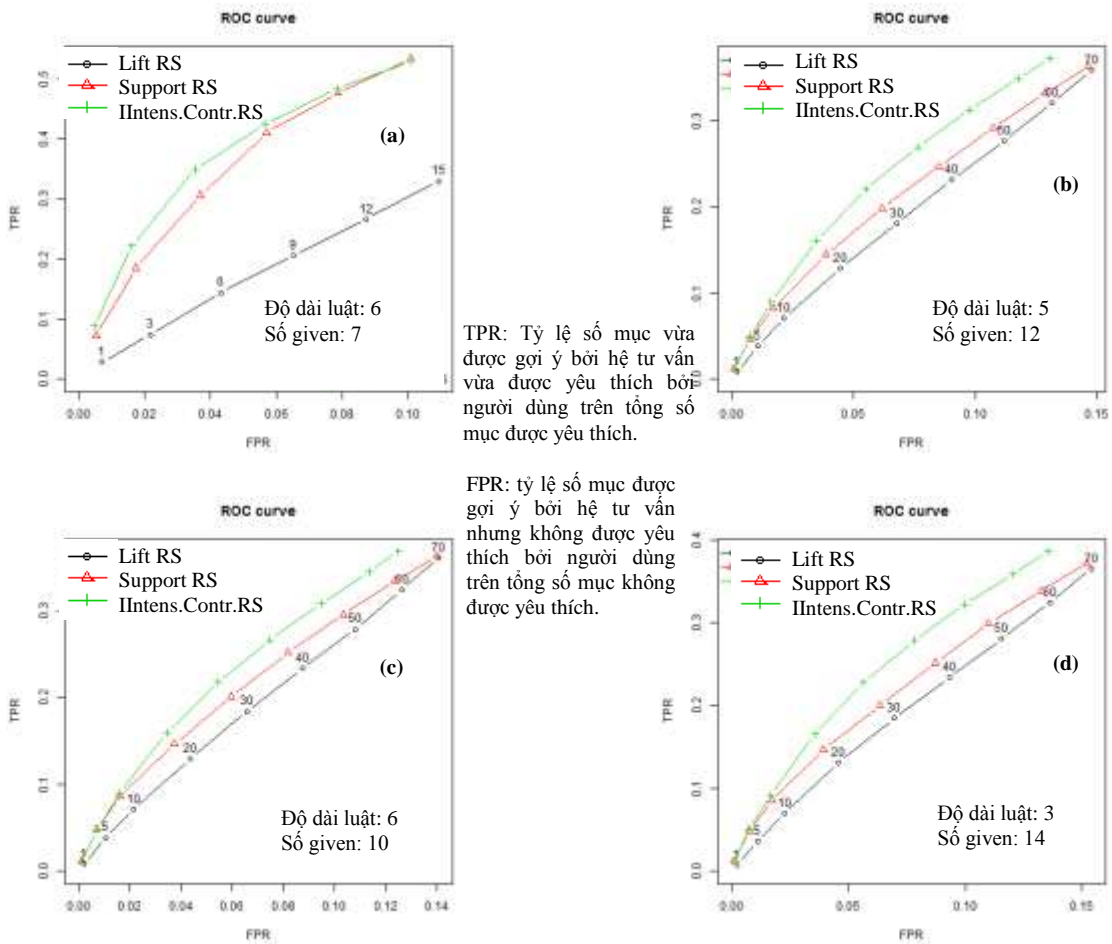


Hình 8. Đường cong precision-recall của ba hệ tư vấn BayesfactorRS, IIntensityRS, IIntens.Contr.RS sử dụng tập dữ liệu MovieLens

E. Kịch bản 4: đánh giá kết quả gợi ý của hệ tư vấn dựa trên mô hình đề xuất với kết quả của hệ tư vấn sử dụng mô hình tư vấn dựa trên luật của recommenderlab

1. Kết quả thực nghiệm

Kịch bản này so sánh các giá trị TPR và FPR của kết quả gợi ý từ ba hệ tư vấn Lift RS, Support RS, IIntens.Contr.RS. Hệ tư vấn Lift RS và Support RS sử dụng mô hình tư vấn dựa trên luật AR được tích hợp trong gói recommenderlab [15]; hệ thứ nhất sử dụng độ đo hỗ trợ support còn hệ thứ hai sử dụng độ đo lift để lọc tập luật và xếp hạng các gợi ý. Hình 9 hiển thị các đường cong ROC của ba hệ thống trong các trường hợp: (độ dài luật tối đa, số given) lần lượt là (6,7) đối với tập dữ liệu MSWeb; (6,10) (5, 12), và (3, 14) đối với tập dữ liệu MovieLens; số mục cần gợi ý cho mỗi người dùng là 1, 3, 6, 9, 12, 15 đối với tập dữ liệu MSWeb và 1, 5, 10, 20, 30, 40, 50, 60, 70 đối với tập dữ liệu MovieLens. Khi thay đổi các trường hợp, ta cũng được kết quả gần tương tự.



Hình 9. Đường cong ROC của ba hệ tư vấn Lift RS, Support RS và IIntens.Contr.RS sử dụng tập dữ liệu MSWeb (Hình 9.a) và MovieLens (Hình 9.b, 9.c, 9.d)

2. Thảo luận

Kết quả thực nghiệm cho thấy đường cong ROC của hệ tư vấn dựa trên mô hình kết hợp cường độ hàm ý với độ đo trách nhiệm IIntens.Contr.RS nằm trên đường cong ROC của hai hệ tư vấn dựa trên mô hình AR sử dụng độ đo support/lift trong recommenderlab (Lift RS và Support RS). Tỷ lệ số mục vừa được gợi ý bởi hệ tư vấn vừa được yêu thích bởi người dùng trên tổng số mục được yêu thích (TPR) của IIntens.Contri.RS cao hơn so với của hai hệ tư vấn còn lại đồng thời tỷ lệ số mục được gợi ý bởi hệ tư vấn nhưng không được yêu thích bởi người dùng trên tổng số mục không được thích (FPR) của IIntens.Contri.RS thấp hơn so với của hai hệ tư vấn đó.

V. KẾT LUẬN

Bài báo này đề xuất một mô hình tư vấn mới dựa trên luật kết hợp và một số độ đo hàm ý thống kê quan trọng. Cụ thể, độ hỗ trợ và độ tin cậy được sử dụng để tạo ra tập luật kết hợp trong khi độ đo cường độ hàm ý kết hợp với độ đo trách nhiệm được sử dụng để lọc tập luật và xếp hạng các gợi ý. Mô hình đề xuất hoàn toàn có thể được dùng để xây dựng các hệ tư vấn dựa trên luật kết hợp nhưng sử dụng những độ đo hấp dẫn khác để lọc tập luật và xếp hạng các gợi ý thông qua bộ bản số $\{n, n_a, n_b, n_{ab}\}$. Các kết quả thực nghiệm cho thấy hệ tư vấn được xây dựng bằng mô hình đề xuất: (1) làm giảm kích thước của tập luật (được sử dụng để đưa ra các gợi ý) trong khi vẫn đảm bảo tính chính xác của sự tư vấn, qua đó làm giảm thời gian thực hiện tư vấn; (2) cho kết quả gợi ý khi kết hợp cường độ hàm ý với độ đo trách nhiệm tốt hơn so với kết quả gợi ý khi sử dụng hệ số bayesian và thống kê chi-square; (3) cho kết quả gợi ý khi kết hợp cường độ hàm ý với độ đo trách nhiệm tốt hơn so với kết quả gợi ý khi sử dụng mô hình tư vấn dựa trên luật AR trong recommenderlab kết hợp với các độ đo support hay lift. Tuy nhiên, để có thể áp dụng mô hình đề xuất này vào việc xây dựng một hệ tư vấn thực tế, mô hình cần được tiếp tục đánh giá và so sánh với một vài mô hình tư vấn phổ biến khác.

TÀI LIỆU THAM KHẢO

- [1] A. Asuncion and D. J. Newman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science. <http://www.ics.uci.edu/~mlearn/MLRepository.html>, 2007.
- [2] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger, “Basic Approaches in Recommendation Systems”, Recommendation Systems in Software Engineering, pp. 15-38, Springer, 2014.
- [3] A. Gunawardana and G. Shani, “A Survey of Accuracy Evaluation Metrics of Recommendation Tasks”, Journal of Machine Learning Research, 10, 2009, pp. 2935–2962.
- [4] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Application of Dimensionality Reduction in Recommender Systems-A Case Study”, Proc. ACM WebKDD Workshop, 2000.
- [5] C. Aggarwal, Recommender Systems: The Textbook, Springer International Publishing Switzerland, 2016.
- [6] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, Recommender Systems Handbook, Springer US, 2011.
- [7] 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, 17, pp. 734-749, 2005.
- [8] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: a survey”, Decision Support Systems, 74, pp. 12-32, 2015.
- [9] L. P. Phan, K. M. Nguyen, H. H. Huynh, and H. X. Huynh, "Association-based recommender system using statistical implicative cohesion measure," 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE), Hanoi, 2016, pp. 144-149. doi: 10.1109/KSE.2016.7758044
- [10] L. P. Phan, N. Q. Phan, K. M. Nguyen, H. H. Huynh, H. X. Huynh, F. Guillet, “Interestingnesslab: A Framework for Developing and Using Objective Interestingness Measures”. Advances in Intelligent Systems and Computing, vol 538. Springer, Cham, pp. 302-311, 2017
- [11] P-N. Tang, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson, 2006.
- [12] R. Gras and P. Kuntz, “An overview of the Statistical Implicative Analysis (SIA) development”, Statistical Implicative Analysis, Studies in Computational Intelligence, Springer-Verlag, 127, 2008, pp. 11-40.
- [13] R. Gras, Chapitre 4: Extension de l’ASI aux variables non binaires, 2013.
- [14] X. Amatriain, A. Jaimes, N. Oliver, and J. Pujol, “Data mining methods for recommender systems”, Recommender Systems Handbook, Springer US, pp. 39-71, 2011.
- [15] recommenderlab, <https://cran.r-project.org/web/packages/recommenderlab/index.html>.

RECOMMENDER SYSTEMS BASED ON IMPLICATIVE INTENSITY AND CONTRIBUTION MEASURES

Lan Phuong Phan, Hung Huu Huynh, Hiep Xuan Huynh

ABSTRACT: *Recommender systems are used to predict the user’s preferences for a specific product or service as well as to recommend the suitable products or services to users. This paper proposes a new recommendation model that uses the support and confidence measures to generate the association rules; the implicative intensity and contribution measures to filter the ruleset and rank the recommendations. The model can be expanded with other interestingness measures. The recommender system based on the proposed model is developed and evaluated by four scenarios and on two datasets. The experimental results show that the recommender system based on implicative intensity and contribution measures can : decrease the size of ruleset but still maintain the accuracy of the recommendation, therefore decrease time to make the recommendation; make the recommendation result that is better than the result of system based on chi-square statistics and bayesian facto; make the recommendation result that is better than the result of systems based on the recommendation model AR (integrated in recommenderlab) and the support/the lift measures.*