

HỆ TƯ VẤN DỰA TRÊN TIẾP CẬN LUẬT KẾT HỢP VÀ ĐỘ ĐO HÀM Ý THỐNG KÊ

Phan Quốc Nghĩa¹, Nguyễn Minh Kỳ², Nguyễn Tấn Hoàng³, 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 Trung tâm Tin học tỉnh Đồng Tháp

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@ctu.edu.vn, hoangntdt@gmail.com, hxhiep@ctu.edu.vn

TÓM TẮT— Trong bài viết này chúng tôi đề xuất một tiếp cận mới nhằm hỗ trợ tư vấn cho thí sinh trung học phổ thông xác định nghề nghiệp phù hợp sẽ làm trong tương lai dựa trên tiếp cận về hệ tư vấn (recommender system). Mô hình hệ tư vấn được xây dựng trên cơ sở phương pháp hàm ý thống kê (statistical implication). Khác với các hệ tư vấn truyền thống chủ yếu tập trung về tính chất logic thể hiện sự tồn tại hay không tồn tại mỗi quan hệ ưu tiên giữa người dùng và mục dữ liệu, trong mô hình này chúng tôi đặc biệt quan tâm đến tỷ lệ hay mối quan hệ hàm ý (implicative) giữa người dùng và mục dữ liệu trong một ngữ cảnh cụ thể. Thông qua thực nghiệm trên hai tập dữ liệu cho thấy rằng mô hình mà chúng tôi đề xuất đưa ra kết quả khuyến nghị khá hữu ích.

Từ khóa— Độ đo hấp dẫn, hàm ý thống kê, định hướng nghề nghiệp, tuyển sinh, hệ tư vấn.

I. GIỚI THIỆU

Hệ tư vấn [2][3][7][5] trở thành lĩnh vực nghiên cứu quan trọng từ khi bắt đầu xuất hiện bài báo đầu tiên về phương pháp lọc cộng tác [15][27][28][29] vào giữa thập niên 1990 của thế kỷ 20. Trong gần hai mươi năm qua, các nghiên cứu về hệ tư vấn đã có được những thành tựu đáng kể trong việc ứng dụng vào cuộc sống cũng như phục vụ nghiên cứu. Hệ tư vấn được hình thành trên cơ sở xác định hay tư vấn một cách có ý nghĩa cho một nhóm người dùng (users) đối với các mục dữ liệu (items) hay sản phẩm (product) mà họ cần quan tâm trên cơ sở xếp hạng các mục dữ liệu theo thứ tự ưu tiên. Các tiếp cận thông dụng được hình thành trên cơ sở nội dung (content-based) [2], lọc cộng tác (collaborative filtering) [2][20] hoặc phối hợp cả hai tiếp cận trên tạo thành tiếp cận hỗn hợp (hybrid) [3][2]. Hiện tại, sự quan tâm nghiên cứu về hệ tư vấn đang được nhiều nhà nghiên cứu quan tâm như: nghiên cứu tìm ra các thuật toán mới để cải thiện độ chính xác của các mô hình hệ tư vấn đã được xuất, nghiên cứu cải tiến các hệ thống để thích ứng với vấn đề bùng nổ thông tin và nghiên cứu để đề xuất các mô hình hệ tư vấn mới.

Kết quả nghiên cứu về hệ tư vấn trong thời gian qua khá đa dạng như các nghiên cứu để ứng dụng các thuật toán nhằm cải tiến độ tin cậy của các mô hình [7][16][17][19][22][23], các nghiên cứu đánh giá mô hình và thuật toán của hệ tư vấn [8][9][14] nghiên cứu phân rã ma trận (matrix factorization) sự tồn tại giá trị giữa người dùng và mục dữ liệu cần quan tâm mang tính chất dự báo sự ưu tiên [13], một số tiếp cận khác quan tâm về tính hành động của mục dữ liệu (actionable) kết hợp với luật kết hợp để tư vấn cho người dùng [14] hay tiếp cận với trừu tượng hóa và luật kết hợp [31][30]. Tuy nhiên, các thể hệ hiện tại của hệ tư vấn vẫn cần cải tiến hơn nữa để có được các phương pháp khuyến nghị hiệu quả hơn và áp dụng với một phạm vi rộng hơn cho các ứng dụng thực tế cuộc sống.

Trong bài báo này, chúng tôi tiếp tục phát triển ý tưởng xây dựng hệ tư vấn dựa trên phân tích hàm ý thống kê [30], mô hình hệ tư vấn dựa trên tiếp cận luật kết hợp và độ đo hàm ý thống kê nhằm khắc phục nhược điểm của các hệ thống tư vấn truyền thống là chủ yếu tập trung về tính chất logic thể hiện sự tồn tại hay không tồn tại mỗi quan hệ ưu tiên giữa người dùng và mục dữ liệu hay sản phẩm. Trong mô hình này chúng tôi đặc biệt quan tâm đến tỷ lệ hay mối quan hệ hàm ý (implicative) giữa người dùng và mục dữ liệu trong một ngữ cảnh cụ thể để đưa ra các khuyến nghị cho người dùng hiệu quả hơn.

Bài viết này được tổ chức thành 6 phần. Phần 1 giới thiệu chung về hệ tư vấn và nêu vấn đề nghiên cứu. Phần 2 giới thiệu các mô hình thông dụng của hệ tư vấn và các nghiên cứu có liên quan. Phần 3 trình bày về độ đo hàm ý thống kê và công thức tính giá trị hấp dẫn dựa trên 4 tham số của hai độ đo hàm ý thống kê. Phần 4 mô tả cách xây mô hình hệ tư vấn dựa trên tiếp cận luật kết hợp và độ đo hàm ý thống kê. Phần 5 trình bày kết quả thực nghiệm của mô hình. Phần cuối cùng tóm tắt một số kết quả quan trọng đã đạt được.

II. HỆ TƯ VẤN

Bài toán tư vấn được coi là bài toán ước lượng trước hạng [2] (rating) của các mục dữ liệu hay các sản phẩm chưa được người dùng xem xét. Việc ước lượng này thường được dựa trên những đánh giá đã có của chính người dùng đó hoặc những người dùng khác. Những mục dữ liệu hay sản phẩm có hạng cao nhất sẽ được dùng để tư vấn. Một cách hình thức, bài toán tư vấn được mô tả như sau:

Gọi C là tập tất cả người dùng, S là tập tất cả các mục dữ liệu hay sản phẩm có thể tư vấn. Tập S có giá trị trong khoảng $\{1, n\}$, tập C có giá trị trong khoảng $\{1, m\}$. Hàm $u(c, s)$ đo độ phù hợp (hay hạng) của mục dữ liệu hay sản phẩm s với người dùng c : $u: C \times S \rightarrow R$ với R là tập được sắp thứ tự. Với mỗi người dùng $c \in C$, cần tìm sản phẩm $s' \in S$ sao cho hàm $u(s', c)$ đạt giá trị lớn nhất [2]:

$$\forall c \in C, s'_c = \arg \max_{s \in S} u(c, s)$$

Trong hệ tư vấn, độ phù hợp của một mục dữ liệu hay sản phẩm thường được cho bằng điểm, ví dụ người dùng A đánh giá bộ phim “Star war 3” là mức 7 điểm trên 10 điểm. Tuy nhiên, cách đo độ phù hợp trong các hệ thống tư vấn có thể là một hàm bất kỳ tùy thuộc vào lĩnh vực ứng dụng cụ thể. Giá trị của hàm u có thể được xác định bởi người dùng hoặc được tính toán bởi công thức nào đó. Mỗi người dùng trong không gian C được xác định bởi một hồ sơ người dùng [1] (User profile). Hồ sơ người dùng này có thể gồm rất nhiều loại thông tin: tên, tuổi, giới tính, thu nhập, vv hoặc có thể chỉ gồm một trường mã số người dùng (user id) duy nhất. Tương tự như thế, mỗi sản phẩm trong không gian S cũng được xác định bởi một tập các đặc trưng. Ví dụ, trong hệ thống tư vấn ngành học, đặc trưng của mỗi ngành có thể là: tên ngành, yêu cầu kiến thức, yêu cầu năng lực, yêu cầu kỹ năng, vv.

Vấn đề chính của hệ tư vấn là hàm u không được xác định trên toàn không gian $C \times S$ mà chỉ trên một miền nhỏ của không gian đó [2]. Điều này dẫn tới việc hàm u phải được ngoại suy trong không gian $C \times S$. Thông thường, độ phù hợp được thể hiện bằng điểm và chỉ xác định trên tập các mục dữ liệu hay sản phẩm đã từng được người dùng đánh giá từ trước thường rất nhỏ. Có rất nhiều cách để dự đoán, ước lượng hạng hoặc điểm cho các mục dữ liệu hay sản phẩm như sử dụng học máy, lý thuyết xấp xỉ, các thuật toán dựa trên kinh nghiệm... Các hệ thống tư vấn thường được phân thành ba loại dựa trên cách nó dùng để ước lượng hạng hoặc điểm của mục dữ liệu hay sản phẩm:

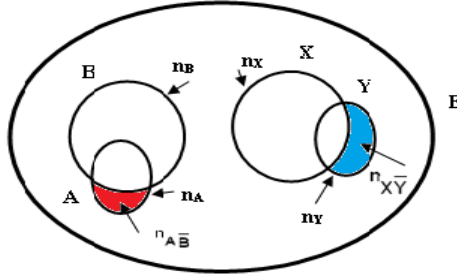
Phương pháp tư vấn dựa trên nội dung (Content-based recommendations), độ phù hợp $u(c, s)$ của mục dữ liệu hay sản phẩm s với người dùng c được đánh giá dựa trên độ phù hợp $u(c, s_i)$, trong đó $s_i \in S$ và “tương tự” như s [2]. Ví dụ, để gợi ý một bộ phim cho người dùng c , hệ thống tư vấn sẽ tìm các đặc điểm của những bộ phim từng được c đánh giá cao (như diễn viên, đạo diễn...), sau đó chỉ những bộ phim tương đồng với sở thích của c mới được giới thiệu.

Không giống như phương pháp tư vấn dựa trên nội dung, hệ thống tư vấn cộng tác (Collaborative recommendations) dự đoán độ phù hợp $u(c, s)$ của một mục dữ liệu hay sản phẩm s với người dùng c dựa trên độ phù hợp $u(c_j, s)$ giữa người dùng c_j và s , trong đó c_j là người có cùng sở thích với c [2][28]. Ví dụ, để gợi ý một bộ phim cho người dùng c , đầu tiên hệ thống cộng tác tìm những người dùng khác có cùng sở thích phim ảnh với c . Sau đó, những bộ phim được họ đánh giá cao sẽ được dùng để tư vấn cho c . Hệ thống cộng tác có thể chia thành hai loại: hệ thống dựa trên kinh nghiệm (heuristic-based hay memory-based) và hệ thống dựa trên mô hình (model-based).

Phương pháp lai ghép (Hybrid approaches), một vài hệ tư vấn kết hợp cả phương pháp cộng tác và dựa trên nội dung nhằm tránh những hạn chế của cả hai phương pháp trên.

III. ĐỘ ĐO HÀM Ý THỐNG KÊ

Hàm ý thống kê (statistical implication) [33][34] nghiên cứu mối quan hệ hàm ý giữa các biến dữ liệu (variable) hay thuộc tính dữ liệu, cho phép phát hiện các luật (rules) $A \rightarrow B$ 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 ”.



Hình 1: Mô hình biểu diễn luật hàm ý thống kê $A \rightarrow B$

Chọn ngẫu nhiên và độc lập hai tập con nào đó X và Y của E (không có mối quan hệ nào giữa X và Y , $X \cap Y = \emptyset$), đồng thời cũng chọn tương tự hai tập con A và B . Cho \bar{Y} và \bar{B} tương ứng là phần bù của Y và B trong E và có cùng bản số $n_{\bar{B}} = n - n_B$. Khi đó luật hàm ý $A \rightarrow B$ sẽ được chấp nhận ở mức tin cậy $1 - \alpha$ nếu và chỉ nếu [34]:

$$\Pr[\text{card}(X \cap \bar{Y}) \leq \text{card}(A \cap \bar{B})] \leq \alpha$$

Và biến $\text{card}(X \cap \bar{Y})$ tuân theo luật phân phối Poisson với tham số $\frac{n_A n_{\bar{B}}}{n}$. Trong trường hợp $n_{\bar{B}} \neq 0$, ta có thể rút gọn và tập trung biến Poisson này thành biến [34]:

$$Q(A, \bar{B}) = \frac{\text{card}(X \cap \bar{Y}) - \frac{n_A(n - n_B)}{n}}{\sqrt{\frac{n_A(n - n_B)}{n}}}$$

Trong thực tế thực nghiệm, giá trị quan sát được của $Q(A, \bar{B})$ là $q(A, \bar{B})$. Chỉ số hàm ý (implicative index) được xác định như sau [34]:

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

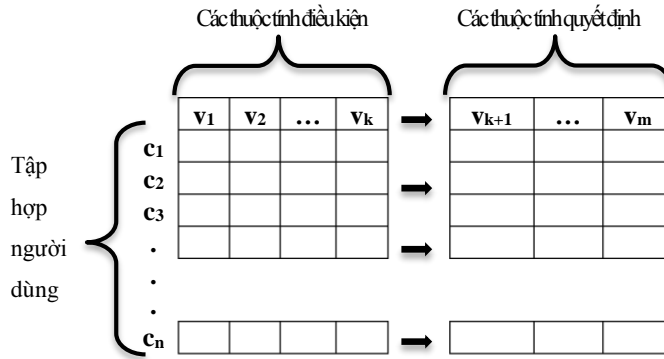
Trong trường hợp xấp xỉ phù hợp (chẳng hạn như $\frac{n_{A\bar{B}}}{n} > 4$), biến $Q(A, \bar{B})$ sẽ có giá trị xấp xỉ theo luật phân phối chuẩn rút gọn. Chỉ số mật độ hàm ý (implication intensity) với $n_A \leq n_B$ và $n_B \neq n$ được định nghĩa trên cơ sở chỉ số $q(A, \bar{B})$ là [34]:

$$\varphi(A, B) = 1 - \Pr[Q(A, \bar{B}) \leq q(A, B)] = \frac{1}{\sqrt{2\pi}} \int_{q(A, \bar{B})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Luật hàm ý $A \rightarrow B$ được chấp nhận ở mức tin cậy $1 - \alpha$ nếu và chỉ nếu $\varphi(A, B) \geq 1 - \alpha$ [34].

IV. HỆ TƯ VẤN THEO HƯỚNG TIẾP CẬN LUẬT KẾT HỢP VÀ ĐỘ ĐO HÀM Ý THỐNG KÊ

Giả sử rằng chúng ta có một tập hợp n người dùng $C = \{c_1, c_2, \dots, c_n\}$ trong một hệ thống bất kỳ. Mỗi người dùng được mô tả bằng một hồ sơ người dùng [1] bao gồm tập hợp các thuộc tính $V = \{v_1, v_2, \dots, v_m\}$, trong đó các thuộc tính từ v_1 đến v_k là thuộc tính điều kiện, các thuộc tính từ v_{k+1} đến v_m là các thuộc tính quyết định. Trên cơ sở hai tập hợp C và V chúng tôi xây dựng mô hình hệ tư vấn dựa trên luật kết hợp và độ đo hàm ý thống kê như sau:



Hình 2: Mô hình hệ tư vấn dựa trên luật kết hợp và độ đo hàm ý thống kê

Trong mô hình chúng tôi xây dựng giải thuật xử lý của hệ thống tư vấn gồm các bước sau:

Giải thuật cho mô hình hệ tư vấn dựa trên luật kết hợp và độ đo hàm ý thống kê

Input: Tập dữ liệu mô tả hồ sơ người dùng gồm n đối tượng $C = \{c_1, c_2, \dots, c_n\}$, mỗi đối tượng được mô tả bằng m thuộc tính $V = \{v_1, v_2, \dots, v_m\}$, trong đó các thuộc tính từ v_1 đến v_k là thuộc tính điều kiện, các thuộc tính từ v_{k+1} đến v_m là các thuộc tính quyết định.

Output: Bảng tính giá trị hấp dẫn cho từng luật kết hợp được sắp xếp dựa trên hai độ đo hàm ý thống kê.

Begin

1: Xử lý tập dữ liệu thô

Chọn tập các thuộc tính theo câu của hệ thống;

For $i=1$ to n

Chọn các dòng dữ liệu tương ứng với tập thuộc tính;

02: Sinh tập luật và lựa chọn các luật cho hệ tư vấn

Sinh tập luật kết hợp từ tập dữ liệu đã xử lý;

Lựa chọn các luật cho hệ tư vấn;

03: Đếm các tham số $n, n_A, n_B, n_{A\bar{B}}$

Với mỗi luật trong tập luật đã chọn

Begin

Đếm tham số n ;

Đếm tham số n_A ;

Đếm tham số n_B ;

Đếm tham số $n_{A\bar{B}}$;

End;

Tạo bảng dữ liệu các tham số cho tập luật;

Luật	n	n_A	n_B	$n_{A\bar{B}}$

04: Áp dụng các độ đo hàm ý thống kê để xếp hạng các luật

Với mỗi dòng tham số $n, n_A, n_B, n_{A\bar{B}}$ trong tập luật đã chọn

Begin

Implicationindex($n, n_A, n_B, n_{A\bar{B}}$);

Implicationintensity($n, n_A, n_B, n_{A\bar{B}}$);

End;

05: Tạo bảng tính giá trị hấp dẫn cho từng luật kết hợp

Kq=

Luật	Implicationindex	Implicationintensity

Return(Kq)

End;

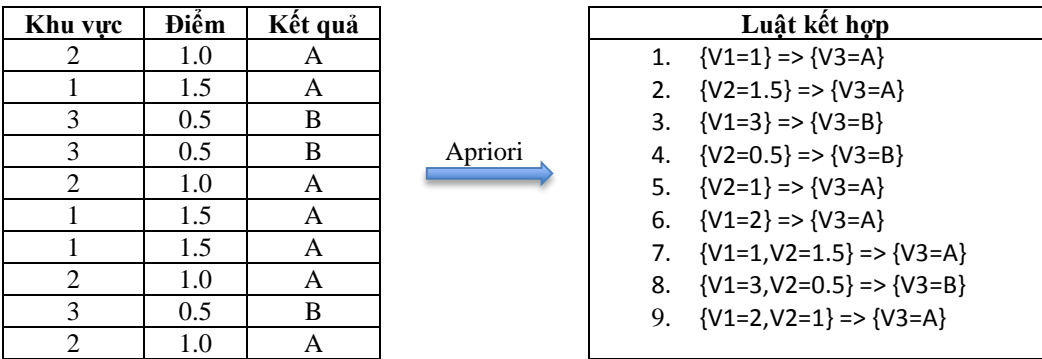
4.1. Xử lý dữ liệu

Trong giai đoạn xử lý dữ liệu, dựa trên hồ sơ người dùng (tập thuộc tính) và các yêu cầu của bài toán tư vấn để xác định tập các thuộc tính điều kiện và tập các thuộc tính quyết định có dạng: $\{v_1, v_2, \dots, v_k, v_{k+1}, \dots, v_m\}$.

4.2. Sinh tập luật và lựa chọn các luật cho hệ tư vấn

Trong bước sinh luật kết hợp, chúng tôi đề xuất áp dụng giải thuật Apriori [35] để sinh ra tập luật kết hợp từ tập dữ liệu đã xử lý. Thông thường tập luật kết hợp được sinh ra có số lượng rất lớn. Vì thế, tùy theo yêu cầu đặt ra của bài toán tư vấn mà chúng ta chọn lựa các luật có dạng $\{v_1, v_2, \dots, v_k\} \rightarrow \{v_{k+1}, \dots, v_m\}$, trong đó: $\{v_1, v_2, \dots, v_k\}$ là tập các thuộc tính điều kiện, $\{v_{k+1}, \dots, v_m\}$ là tập các thuộc tính quyết định.

Ví dụ 1: Sinh luật từ tập dữ liệu có 3 thuộc tính, trong đó thuộc tính 1, 2 là thuộc tính điều kiện, thuộc tính 3 là thuộc tính quyết định. Áp dụng thuật toán Apriori với Supp=0.01 và conf=0.5 ta sinh được tập luật kết hợp gồm 26 luật, sau đó ta tiến hành lựa chọn các luật có vẻ phải chứa thuộc tính quyết định ta được tập luật kết quả gồm 9 luật được trình bày trong hình 3.

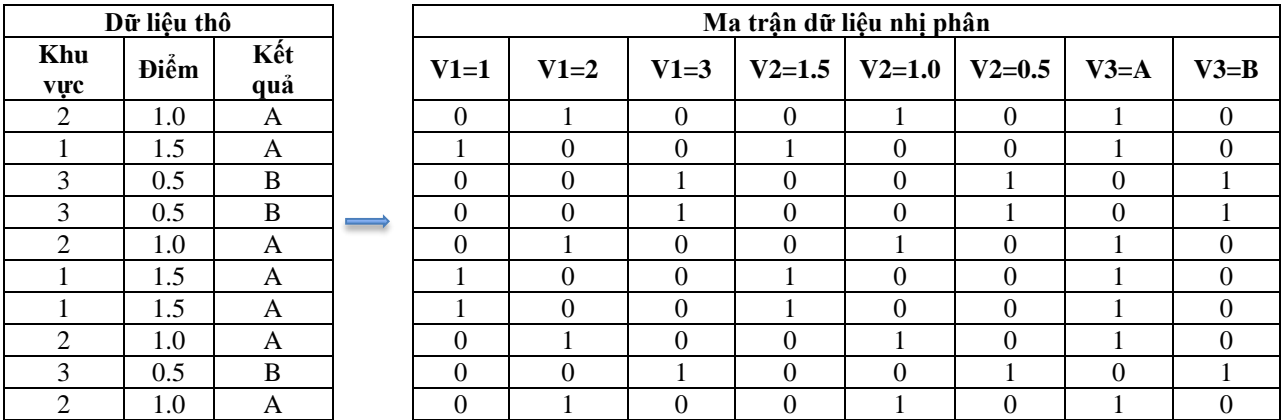


Hình 3: Kết quả sinh luật theo giải thuật Apriori [35] trên tập dữ liệu thô [35] bằng công cụ arules [36]

4.3. Đếm các tham số n, n_A, n_B, n_{AB} của hàm tính độ đo hàm ý

Để áp dụng được các độ đo hàm ý thống kê cũng như các độ đo hấp dẫn khách quan khác thì việc đếm các tham số n, n_A, n_B, n_{AB} của từng luật kết hợp là một bước cần thiết trong quy trình xử lý của hệ thống. Để đáp ứng yêu cầu này, chúng tôi đề xuất phương pháp chuyển đổi tập dữ liệu thô và tập luật kết hợp sang dạng ma trận nhị phân. Đối với tập dữ liệu thô, dựa trên miền giá trị của từng thuộc tính của tập dữ liệu, việc chuyển đổi được thực hiện theo nguyên tắc cứ mỗi giá trị của thuộc tính trên tập dữ liệu thô sẽ sinh ra một cột trên ma trận dữ liệu nhị phân.

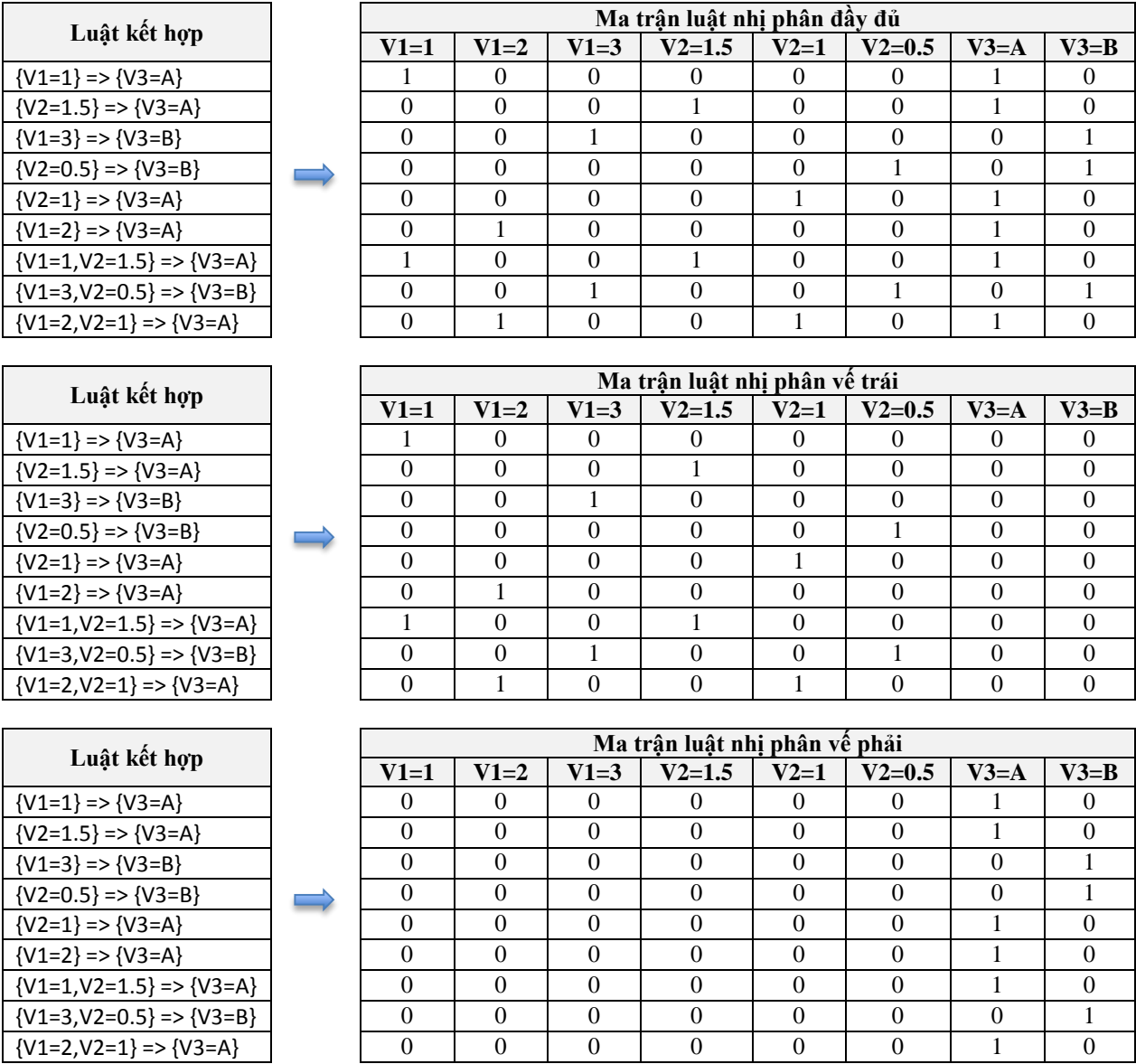
Ví dụ 2: Chuyển đổi dữ liệu thô sang dạng Ma trận dữ liệu nhị phân được thể hiện ở hình 4.



Hình 4: Minh họa chuyển từ dữ liệu thô sang Ma trận dữ liệu nhị phân

Đối với tập luật kết hợp, chúng tôi tách ra thành ba ma trận dạng nhị phân: Ma trận luật nhị phân đầy đủ, Ma trận luật nhị phân về trái và Ma trận luật nhị phân về phải. Trong từng ma trận nhị phân, việc chuyển đổi được thực hiện theo nguyên tắc cứ mỗi thuộc tính xuất hiện trong về trái hoặc về phải của luật kết hợp sẽ sinh ra một cột mới trong ma trận luật nhị phân tương ứng.

Ví dụ 3: Từ tập luật kết hợp ở trên ta chuyển sang ma trận nhị phân tương ứng được thể hiện trong hình 5.



Hình 5: Minh họa chuyển từ luật kết hợp sang Ma trận dữ liệu nhị phân

Sau khi có Ma trận dữ liệu nhị phân, Ma trận luật nhị phân về trái và Ma trận luật nhị phân về phải, bước tiếp theo là tiến hành đếm các tham số $n, n_A, n_B, n_{A\bar{B}}$ cho từng luật trong tập luật đã chọn theo công thức sau:

- Đếm tham số n : n = số dòng của Ma trận dữ liệu nhị phân.
- Đếm tham số n_A : lấy từng dòng của Ma trận luật nhị phân về trái so khớp với tất cả các dòng của Ma trận dữ liệu nhị phân, nếu các bits khớp với dòng luật đang so sánh thì tăng n_A lên 1.
- Đếm tham số n_B : lấy từng dòng của Ma trận luật nhị phân về phải so khớp với tất cả các dòng của Ma trận dữ liệu nhị phân, nếu các bits khớp với dòng luật đang so sánh thì tăng n_B lên 1.
- Đếm tham số $n_{A\bar{B}}$: lấy từng dòng của Ma trận luật nhị phân đầy đủ so khớp với tất cả các dòng của Ma trận dữ liệu nhị phân, nếu các bits khớp với dòng luật đang so sánh thì tăng $n_{A\bar{B}}$ lên 1. Khi đó $n_{A\bar{B}} = n_A - n_{AB}$.

Ví dụ 3: Đếm các tham số $n, n_A, n_B, n_{A\bar{B}}$ từ tập luật kết hợp đã chọn được thể hiện trong bảng 1.

Bảng 1: Kết quả đếm các tham số $n, n_A, n_B, n_{A\bar{B}}$ cho từng luật kết hợp

Luật kết hợp	n	n_A	n_B	$n_{A\bar{B}}$
$\{V1=1\} \Rightarrow \{V3=A\}$	10	3	7	0
$\{V2=1.5\} \Rightarrow \{V3=A\}$	10	2	7	0
$\{V1=3\} \Rightarrow \{V3=B\}$	10	3	3	0
$\{V2=0.5\} \Rightarrow \{V3=B\}$	10	3	3	0
$\{V2=1\} \Rightarrow \{V3=A\}$	10	4	7	0
$\{V1=2\} \Rightarrow \{V3=A\}$	10	4	7	0
$\{V1=1,V2=1.5\} \Rightarrow \{V3=A\}$	10	3	7	0

$\{V1=3,V2=0.5\} \Rightarrow \{V3=B\}$	10	3	3	0
$\{V1=2,V2=1\} \Rightarrow \{V3=A\}$	10	4	7	0

4.3. Áp dụng các độ đo hàm ý thống kê để xếp hạng các luật phục vụ tư vấn

Từ bảng kết quả đếm các tham số n, n_A, n_B, n_{AB} của từng luật ta áp dụng cho hai độ đo hàm ý thống kê: Implication index [34]Error! Reference source not found. và Implication intensity [34]Error! Reference source not found., sau đó xếp hạng theo giá trị độ đo ta được kết quả trong bảng 2.

Bảng 2: Bảng tính giá trị hấp dẫn cho từng luật kết hợp

Luật kết hợp	Implication index	Implication intensity
$\{V2=1.5\} \Rightarrow \{V3=A\}$	-0.7745967	0.780711
$\{V1=1\} \Rightarrow \{V3=A\}$	-0.9486833	0.8286091
$\{V1=1,V2=1.5\} \Rightarrow \{V3=A\}$	-0.9486833	0.8286091
$\{V2=1\} \Rightarrow \{V3=A\}$	-1.095445	0.8633392
$\{V1=2\} \Rightarrow \{V3=A\}$	-1.095445	0.8633392
$\{V1=2,V2=1\} \Rightarrow \{V3=A\}$	-1.095445	0.8633392
$\{V1=3\} \Rightarrow \{V3=B\}$	-1.449138	0.9263504
$\{V2=0.5\} \Rightarrow \{V3=B\}$	-1.449138	0.9263504
$\{V1=3,V2=0.5\} \Rightarrow \{V3=B\}$	-1.449138	0.9263504

Từ kết quả này, khi cần tư vấn cho một đối tượng mới, chúng ta dựa trên các thuộc tính điều kiện của đối tượng này và kết quả xếp hạng của tập luật dựa trên các độ đo hàm ý thống kê để đưa ra các khuyến nghị cho đối tượng mới lựa chọn các giá trị của thuộc tính quyết định một cách hợp lý nhất.

V. KẾT QUẢ THỰC NGHIỆM

5.1. Dữ liệu sử dụng

Trong phần thực nghiệm này, để kiểm tra độ chính xác và khả năng ứng dụng của mô hình, 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 Lenses của UCI [37]. Kịch bản 2 chúng tôi sử dụng tập dữ liệu tuyển sinh đại học, cao đẳng của Trường Đại học Trà Vinh [38].

Trong kịch bản 1, chúng tôi tiến hành thực nghiệm trên tập dữ liệu Lenses (Cơ sở dữ liệu cho việc chọn kính sát trùng) của UCI [37]. Tập dữ liệu này gồm 24 mẫu tin, 5 thuộc tính (trong đó 4 thuộc tính đầu là thuộc tính điều kiện và thuộc tính thứ 5 là thuộc tính quyết định. Nội dung của tập dữ liệu được trình bày trong bảng 3.

Bảng 3: Tập dữ liệu Lenses của UCI

TT	V1	V2	V3	V4	V5	TT	V1	V2	V3	V4	V5
1	1	1	1	1	3	13	2	2	1	1	3
2	1	1	1	2	2	14	2	2	1	2	2
3	1	1	2	1	3	15	2	2	2	1	3
4	1	1	2	2	1	16	2	2	2	2	3
5	1	2	1	1	3	17	3	1	1	1	3
6	1	2	1	2	2	18	3	1	1	2	3
7	1	2	2	1	3	19	3	1	2	1	3
8	1	2	2	2	1	20	3	1	2	2	1
9	2	1	1	1	3	21	3	2	1	1	3
10	2	1	1	2	2	22	3	2	1	2	2
11	2	1	2	1	3	23	3	2	2	1	3
12	2	1	2	2	1	24	3	2	2	2	3

Trong kịch bản 2, chúng tôi tiến hành thực nghiệm trên tập dữ liệu tuyển sinh của Trường Đại học Trà Vinh [38]. Đây là tập dữ liệu được xuất ra từ phần mềm tuyển sinh của Trường với tổng số mẫu tin là 12287 trong đó có 3202 mẫu tin của năm 2012, 3624 mẫu tin của năm 2013 và 5461 mẫu tin của năm 2014. Trong tập dữ liệu này mỗi học sinh được mô tả bằng một hồ sơ gồm 71 thuộc tính (71 cột). Đây là dữ liệu tuyển sinh đặt trưng cho phương thức tuyển sinh theo Quy chế thi của Kỳ thi tuyển sinh ba chung. Trong kỳ thi này, học sinh đã làm hồ sơ dự thi để chọn Trường, chọn Ngành trước khi tham gia kỳ thi.

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 RStudio. Đây là gói công cụ được nhóm chúng tôi phát triển từ nền tảng của công cụ ARQAT phát triển trên Java [39]. Trong gói công cụ này, chúng tôi cập nhật khá đầy đủ các hàm tính độ đo hấp dẫn khách quan cho luật kết hợp dựa trên 4 tham số n, n_A, n_B, n_{AB} , các chức năng xử lý dữ liệu, tích hợp hàm sinh luật Apriori của gói arules [36], chức năng đếm các tham số, n, n_A, n_B, n_{AB} , chức năng tính độ hấp dẫn của luật kết hợp theo các độ đo được chọn, chức xếp hạng luật theo giá trị hấp dẫn, chức năng sinh kết quả tư vấn.

5.3. Kịch bản 1

Với mục tiêu kiểm tra độ chính xác của mô hình, chúng tôi tiến hành chạy từng bước của mô hình trên tập dữ liệu chuẩn Lenses do UCI công bố. Kết quả sinh luật của hàm Apriori với Supp=0.15 và conf=0.8 là 23 luật kết hợp. Từ kết quả này, chúng tôi chọn các luật có vẻ phải chứa thuộc tính quyết định cho hệ thống tư vấn gồm 10 luật được trình bày trong bảng 4.

Bảng 4: Tập luật kết hợp được sinh ra từ tập dữ liệu Lenses

Luật kết hợp
{V4=1} => {V5=3}
{V3=1,V4=2} => {V5=2}
{V1=1,V4=1} => {V5=3}
{V1=2, V4=1} => {V5=3}
{V1=3,V4=1} => {V5=3}
{V2=1,V4=1} => {V5=3}
{V3=1,V4=1} => {V5=3}
{V3=2, V4=1} => {V5=3}
{V2=2,V4=1} => {V5=3}
{V2=2,V3=2} => {V5=3}

Bước thực nghiệm tiếp theo là đếm các tham số n, n_A, n_B, n_{AB} bằng cách gọi hàm đếm các tham số trong gói công cụ ARQAT **Error! Reference source not found.** cho tập luật đã chọn và tạo ra bảng dữ liệu các tham số của tập luật như trong bảng 5.

Bảng 5: Kết quả đếm các tham số n, n_A, n_B, n_{AB} cho từng luật kết hợp của tập dữ liệu Lenses

Luật kết hợp	n	n_A	n_B	n_{AB}
{V4=1} => {V5=3}	24	12	15	0
{V3=1,V4=2} => {V5=2}	24	6	5	1
{V1=1,V4=1} => {V5=3}	24	4	15	0
{V1=2, V4=1} => {V5=3}	24	4	15	0
{V1=3,V4=1} => {V5=3}	24	4	15	0
{V2=1,V4=1} => {V5=3}	24	6	15	0
{V3=1,V4=1} => {V5=3}	24	6	15	0
{V3=2, V4=1} => {V5=3}	24	6	15	0
{V2=2,V4=1} => {V5=3}	24	6	15	0
{V2=2,V3=2} => {V5=3}	24	6	15	1

Từ bảng dữ liệu các tham số, chúng tôi tiến hành khảo sát kết quả của hai độ đo: Độ đo chỉ số hàm ý (Implication Index) và Độ đo mật độ hàm ý (Implication Intensity) cho mỗi luật được chọn. Từ kết quả khảo sát của hai độ đo trên, chúng tôi tạo ra Bảng tính giá trị hấp dẫn cho từng luật kết hợp đồng thời xếp hạng các luật theo thứ tự giảm dần dựa trên giá trị của hai độ đo hàm ý thống kê. Kết quả của xếp hạng được thể hiện trong bảng 6.

Bảng 6: Bảng tính giá trị hấp dẫn cho từng luật kết hợp của tập dữ liệu Lenses

Luật kết hợp	Implication index	Implication intensity
{V2=2,V3=2} => {V5=3}	-0.8333333333	0.797671619
{V1=1,V4=1} => {V5=3}	-1.224744871	0.889664319
{V1=2,V4=1} => {V5=3}	-1.224744871	0.889664319
{V1=3,V4=1} => {V5=3}	-1.224744871	0.889664319
{V2=1,V4=1} => {V5=3}	-1.5	0.933192799
{V3=1,V4=1} => {V5=3}	-1.5	0.933192799
{V3=2,V4=1} => {V5=3}	-1.5	0.933192799
{V2=2,V4=1} => {V5=3}	-1.5	0.933192799
{V3=1,V4=2} => {V5=2}	-1.720618004	0.957339918
{V4=1} => {V5=3}	-2.121320344	0.983052573

Từ bảng tính giá trị hấp dẫn cho từng luật kết hợp này, chúng tôi gọi hàm sinh kết quả tư vấn với dữ liệu đầu vào là các thuộc tính điều kiện và kết quả đầu ra là tập các giá trị của thuộc tính quyết định được xếp theo thứ tự ưu tiên. Với các giá trị của thuộc tính quyết định này, hệ thống sẽ đưa ra các khuyến nghị cho người dùng khi cần chọn mua một kính sát trùng theo các thuộc tính điều kiện đầu vào đã cho trước.

5.4. Kịch bản 2

Với mục tiêu kiểm tra khả năng ứng dụng vào thực tế của mô hình và mong muốn xây dựng hệ tư vấn chọn ngành học cho học sinh, chúng tôi chọn tập dữ liệu tuyển sinh đại học, cao đẳng của Trường Đại học Trà Vinh để chạy thử mô hình lần thứ 2. Sau khi xem xét yêu cầu bài toán tư vấn chọn ngành học cho học sinh khi biết điểm thi các môn trong kỳ

thi THPT quốc gia hoặc kết quả học tập các môn của ba năm học trung học phổ thông, chúng tôi tiến hành xử lý dữ liệu cho bài toán trên công cụ ARQAT **Error! Reference source not found.**cho kết quả trong bảng 7.

Bảng 7: Kết quả xử lý dữ liệu tuyển sinh trên công cụ ARQAT

row.names	V1	V2	V3	V4	V5	V6	V7
1	A	1	1	300	275	300	D140201
2	A	1	1	350	175	350	D140201
3	D1	1	1	175	300	425	D140201
4	D1	1	1	225	200	475	D140201
5	D1	1	1	150	400	475	D140201
6	C	1	1	650	300	450	D140201
7	C	1	1	600	475	550	D140201
8	C	1	1	675	575	550	D140201
9	B	1	1	150	400	325	D720101
10	B	0	2	575	550	500	D720101

Trong bảng 7, các cột từ V1 đến V6 là các cột thuộc tính điều kiện. V1 chứa giá trị của thuộc tính khối thi dùng để xét tuyển vào đại học, cao đẳng của học sinh. Thuộc tính này phụ thuộc vào thông tin tuyển sinh hàng năm của từng trường, từng ngành học. V2 chứa giá trị của thuộc tính giới tính của học sinh. Trong quy chế tuyển sinh không có quy định cụ thể về giới tính khi chọn ngành học nhưng trong thực tế thì mối quan hệ giữa giới tính và ngành học là khá lớn. V3 chứa giá trị của thuộc tính khu vực của học sinh cư trú. Thuộc tính này cho chúng ta biết thông tin về nơi học trung học phổ thông của các em. V4, V5, V6 chứa giá trị của thuộc tính điểm thi của ba môn thi thuộc khối thi. Các thuộc tính này phản ánh trực tiếp năng lực học tập của học sinh, đồng thời cũng là ba thuộc tính điều kiện quan trọng để quyết định chọn ngành học của học sinh bởi vì theo Quy chế tuyển sinh thì để vào học đại học, cao đẳng một ngành nào đó thì học sinh phải có điểm ba môn này đạt trên ngưỡng đảm bảo chất lượng đầu vào hàng năm do Bộ Giáo dục và Đào tạo xác định. V7 chứa giá trị của thuộc tính ngành học. Đây là thuộc tính quyết định của hệ thống tư vấn.

Sau khi xử lý dữ liệu thô chúng tôi tiến hành sinh tập luật kết hợp. Để thu được tập luật kết hợp tương đối đầy đủ, chúng tôi chọn giải pháp sinh luật kết hợp trên dữ liệu của từng năm và khi gọi hàm Apriori chúng tôi chọn độ hỗ trợ và độ tin cậy khá thấp (Supp=0.01, conf=0.5). Xuất phát từ yêu cầu của bài toán là trước khi làm thủ tục xét tuyển vào các trường đại học, cao đẳng học sinh đã có đầy đủ giá trị của các thuộc tính điều kiện (khối thi, giới tính, khu vực, điểm các môn) nhưng chưa chọn được giá trị của thuộc tính quyết định (ngành học), chúng tôi quyết định chọn các luật kết hợp có vẻ phải chứa thuộc tính quyết định (V7) cho hệ thống tư vấn. Kết quả của các bước được trình bày trong bảng 8.

Bảng 8: Kết quả sinh tập luật kết hợp và chọn các luật cho hệ thống trên tập dữ liệu tuyển sinh

Dữ liệu	Số mẫu tin	Số luật sinh ra	Số luật được chọn
Năm 2012	3202	54861	6448
Năm 2013	3624	44801	6118
Năm 2014	5461	53655	6820
Tổng cộng	12287	153317	19386

Với mục tiêu đo độ hấp dẫn các luật dựa trên hàm 4 tham số $f(n, n_A, n_B, n_{AB})$, bước thực nghiệm tiếp theo là đếm các tham số n, n_A, n_B, n_{AB} bằng cách gọi hàm trong công cụ ARQAT **Error! Reference source not found.**cho tập luật đã chọn và tạo ra bảng dữ liệu các tham số của tập luật có cấu trúc như trong bảng 9.

Bảng 9: Kết quả đếm các tham số n, n_A, n_B, n_{AB} cho từng luật kết hợp trên tập dữ liệu tuyển sinh

	Rules	n	nA	nB	nAB_
1	{V4=875} => {V7=D380101}	3624	4	452	0
2	{V6=750} => {V7=D720101}	3624	8	408	4
3	{V6=100} => {V7=D380101}	3624	9	452	4
4	{V5=1000} => {V7=D140201}	3624	9	428	5
5	{V6=675} => {V7=D720101}	3624	9	408	5
6	{V5=750} => {V7=D720101}	3624	10	408	6
7	{V6=125} => {V7=D380101}	3624	13	452	6
8	{V4=775} => {V7=D720101}	3624	14	408	10

9	$\{V3=3\} \Rightarrow \{V7=D720332\}$	3624	15	398	11
10	$\{V4=800\} \Rightarrow \{V7=D720101\}$	3624	16	408	12

Với bảng dữ liệu các tham số này, chúng tôi tiến hành khảo sát độ hấp dẫn của các luật kết hợp đã chọn trên hai độ đo: Độ đo chỉ số hàm ý (Implication Index) và Độ đo mật độ hàm ý (Implication Intensity). Từ kết quả khảo sát của hai độ đo hàm ý thống kê trên, chúng tôi tạo ra Bảng tính giá trị hấp dẫn cho từng luật kết hợp đồng thời xếp hạng các luật theo thứ tự giảm dần dựa trên giá trị của hai độ đo hàm ý thống kê. Kết quả của xếp hạng được thể hiện trong bảng 10.

Bảng 10: Bảng tính giá trị hấp dẫn cho từng luật kết hợp trên tập dữ liệu tuyển sinh

	Rules	ImplicationIndex	ImplicationIntensity
1	$\{V2=1\} \Rightarrow \{V7=D380101\}$	1.060453316	0.144469209
2	$\{V2=1, V3=1\} \Rightarrow \{V7=D380101\}$	0.92445979	0.17762348
3	$\{V5=250\} \Rightarrow \{V7=D720501\}$	0.499585979	0.308683316
4	$\{V2=1, V3=2\} \Rightarrow \{V7=D380101\}$	0.496455457	0.309786553
5	$\{V1=A, V2=1\} \Rightarrow \{V7=D380101\}$	0.466408537	0.320461551
6	$\{V3=1, V5=350\} \Rightarrow \{V7=D380101\}$	0.406170576	0.342308636
7	$\{V1=A, V4=150\} \Rightarrow \{V7=D380101\}$	0.391975268	0.347538244
8	$\{V6=275\} \Rightarrow \{V7=D380101\}$	0.383860603	0.350540884
9	$\{V2=1, V3=1, V5=350\} \Rightarrow \{V7=D380101\}$	0.37465666	0.353957914
10	$\{V2=0\} \Rightarrow \{V7=D720332\}$	0.374603272	0.353977769

Từ kết quả xếp hạng luật theo giá trị hấp dẫn của hai độ đo hàm ý thống kê, chúng tôi xây dựng hệ tư vấn hỗ trợ học sinh chọn ngành học trước khi đăng ký xét tuyển vào các trường đại học, cao đẳng. Trong hệ thống này, các em học sinh sẽ nhập vào hệ thống các giá trị thuộc tính điều kiện như: Khối thi (V1), Phái (V2), Khu vực (V3), Điểm môn 1 (V4), Điểm môn 2 (V5), Điểm môn 3 (V6) (Điểm thi ở đây có thể là điểm thi của Kỳ thi THPT quốc gia hoặc kết quả học tập các môn của ba năm học trung học phổ thông). Hệ thống sẽ dựa trên kết quả xếp hạng luật theo giá trị hấp dẫn của hai độ đo hàm ý thống kê để đưa ra danh mục các ngành theo thứ tự ưu tiên nhằm hỗ trợ các em chọn được ngành học phù hợp với năng lực bản thân các em.

VI. KẾT LUẬN

Trong viết này, chúng tôi đã xây dựng mô hình hệ tư vấn dựa trên tiếp cận luật kết hợp và độ đo hàm ý thống kê. Trong mô hình này, chúng tôi đặc biệt quan tâm đến mối quan hệ hàm ý giữa tập thuộc tính điều kiện và tập thuộc tính quyết định trong từng ngữ cảnh cụ thể. Dựa trên tập luật được sinh ra từ tập dữ liệu mô hình cho phép lựa chọn các luật theo yêu cầu của từng bài toán tư vấn. Khi đã chọn được tập luật cho bài toán, mô hình sẽ tiến hành khảo sát giá trị hấp dẫn của từng luật bằng hai độ đo hàm ý thống kê. Dựa trên kết quả của hai độ đo này, hàm tư vấn sẽ đưa ra kết quả khuyến nghị cho người dùng. Một điểm mới khác biệt giữa mô hình của chúng tôi so với các mô hình hệ tư vấn truyền thống là các mô hình truyền thống thường dựa trên hai đối tượng chính là người dùng và sản phẩm (hay mục dữ liệu) để phát triển các giải thuật tư vấn còn trong mô hình của chúng tôi thì các thuật toán tư vấn được phát triển trên tập thuộc tính của một đối tượng người dùng. Trong các thuật toán này, chúng tôi dựa vào giá trị của các thuộc tính điều kiện để tính toán ra các giá trị của các thuộc tính quyết định của cùng một đối tượng. Qua kết quả thực nghiệm, chúng tôi thấy rằng mô hình của chúng tôi cho kết quả khuyến nghị khá phù hợp với yêu cầu thực tế trên tập dữ liệu chuẩn của UCI đồng thời có thể áp dụng cho bài toán tư vấn chọn ngành học theo yêu cầu của quy định tuyển sinh đại học, cao đẳng hiện nay.

VII. TÀI LIỆU THAM KHẢO

[1] G. Adomavicius and A. Tuzhilin, Expert-Driven Validation of Rule-Based User Models in Personalization Applications, Data Mining and Knowledge Discovery archive, volume 5 Issue 1-2, pp.33-58, 2001.

[2] G. Adomavicius and A. Tuzhilin, Towards 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(6), pp.734-749, 2005.

[3] G. Popescu, Group recommender systems as a voting problem, <http://hci.epfl.ch/projects/GroupFun/>, pp.1-44, 2011.

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

[5] Francesco Ricci, Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer-Verlag and Business Media LLC, pp.1-35, 2011.

[6] Gediminas Adomavicius, Jingjing Zhang, Stability of Recommendation Algorithms, ACM Transactions on Information Systems 30(4), pp.1-31, 2012.

- [7] Guy Shani, David Heckerman, Ronen I. Brafman, An MDP-Based Recommender System, *Journal of Machine Learning Research* (6), pp.1265-1295, 2005.
- [8] Guy Shani and Asela Gunawardana, Evaluating Recommendation Systems, Microsoft Research, pp.1-42, 2009.
- [9] Alan Said, Domonkos Tikk, Andreas Hotho, The Challenge of Recommender Systems Challenges (tutorial), *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.1-2, 2012.
- [10] Xavier Amatriain, Building Industrial-scale Real-world Recommender Systems (tutorial), *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.1-2, 2012.
- [11] Maria Augusta S. N. Nunes, Rong Hu, Personality-based Recommender Systems: An Overview (tutorial), *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.1-2, 2012.
- [12] Bart P. Knijnenburg, Conducting User Experiments in Recommender Systems (tutorial), *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.1-2, 2012.
- [13] Yehuda Koren, Robert Bell and Chris Volinsky, Matrix factorization techniques for recommender systems, *Journal of Computer* 42(8), IEEE Computer Society Press, pp.30-37, 2009.
- [14] Thomas Piton, Julien Blanchard, Fabrice Guillet: CAPRE: A New Methodology for Product Recommendation Based on Customer Actionability and Profitability. *ICDM Workshops 2011*, pp.466-473, 2011.
- [15] Gabor Takacs et al., Scalable collaborative filtering approaches for large recommender systems, *Journal of Machine Learning Research*, pp.623-656, 2009.
- [16] Mario Rodríguez, Christian Posse, Ethan Zhang: Multiple objective optimization in recommender systems, *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.11-18, 2012.
- [17] Xiwang Yang, Harald Steck, Yong Liu: Circle-based recommendation in online social networks, *Proceedings of 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1267-1275, 2012.
- [18] Xiwang Yang, Harald Steck, Yang Guo, Yong Liu: On top-k recommendation using social networks. *ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems*, pp.67-74, 2012.
- [19] R. Bell, Y. Koren, and C. Volinsky, Modeling relationships at multiple scales to improve accuracy of large recommender systems, *Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.95-104, 2007.
- [20] 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.
- [21] P. Bedi, H. Kaur, and S. Marwaha, Trust based recommender system for semantic web, *IJCAI'07 - Proceedings of the 2007 International Joint Conferences on Artificial Intelligence*, pp.2677-2682, 2007.
- [22] H. Ma et al., Improving recommender systems by incorporating social contextual information, *ACM Transactions on Information Systems* 29(2/9), pp.1-23, 2011.
- [23] Christos Zigkolis, Savvas Karagiannidis and Athena Vakali, Dissimilarity Features in Recommender Systems, *ICDMW, 2013, 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW) 2013, pp.825-832, 2013.
- [24] Ferdaous Hdioud, Bouchra Frikh and Brahim Ouhbi, Multi-Criteria Recommender Systems based on MultiAttribute Decision Making, *IWAS '13 Proceedings of International Conference on Information Integration and Web-based Applications & Services*, pp 203, 2013.
- [25] Ting Yuan, Jian Cheng, Xi Zhang, Shuang Qiu, Hanqing Lu, Recommendation by Mining Multiple User Behaviors with Group Sparsity, *AAAI Publications, Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp.222-228, 2014.
- [26] Zhi Qiao, Peng Zhang, Yanan Cao, Chuan Zhou, Li Guo, Bingxing Fang, Combining Heterogenous Social and Geographical Information for Event Recommendation, *AAAI Publications, Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp.145-151, 2014.
- [27] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, Recommending and Evaluating Choices in a Virtual Community of Use, *Human Factors in Computing Systems*, pp.194-201, 1995.
- [28] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, Chapel Hill, pp.175-186, 1994.
- [29] U. Shardanand and P. Maes, Social Information Filtering: Algorithms for Automating 'Word of Mouth', *CHI '95 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 210-217, 1995.
- [30] Raphaël Couturier, Un système de recommandation basé sur l'analyse statistique implicative, *Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference*, pp.151-156, 2005.
- [31] Thu Minh Nguyen Tran, François Sempé, Vinh Tuong Ho, Jean-Daniel Zucker, Mining Abstract Highly Correlated Pairs, *RIVF'09 - 2009 IEEE-RIVF International Conference on Computing and Communication Technologies*, pp.1-4, 2009.
- [32] Thu Minh Nguyen Tran, François Sempé, Jean-Daniel Zucker, AbsTop-Ka : un algorithme d'extraction de paires abstraites hautement corrélées pour mieux recommander dans "la longue traîne", *EGC'10 Extraction et gestion des connaissances (EGC'2010)*, *Revue des Nouvelles Technologies de l'Information (RNTI-E-19)*, Cépaduès-Éditions, ISBN 978-2-85428-922-0, pp.689-690, 2010.
- [33] Régis Gras *et al.*, L'implication statistique – Nouvelle méthode exploratoire de données, La pensée sauvage édition, 1996.

- [34] 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.
- [35] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, pp.487-499, 1994.
- [36] Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik, Mining Association Rules and Frequent Itemsets, Package ‘arules’ <http://R-Forge.R-project.org/projects/arules/>, <http://lyle.smu.edu/IDA/arules/>, 2014.
- [37] UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 1990, [<https://archive.ics.uci.edu/ml/machine-learning-databases/lenses/>]
- [38] Phòng Khảo thí, Trường Đại học Trà Vinh, Dữ liệu tuyển sinh đại học, cao đẳng, 2014.
- [39] Huỳnh Xuân Hiệp, Công cụ ARQAT trên Java, 2005 (ASMDA 2005), <http://conferences.telecom-bretagne.eu/asmda2005/IMG/pdf/proceedings/proceedings-asmda2005.pdf>