Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR); Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.000110

TƯ VẤN DỰA TRÊN ĐỘ BIẾN THIÊN CỦA CHỈ SỐ HÀM Ý TRONG TRƯỜNG HÀM Ý THỐNG KÊ

Nguyễn Tấn Hoàng [1], Huỳnh Hữu Hưng[2], Huỳnh Xuân Hiệp[3]

[1] Sở Thông tin và Truyền thông tỉnh Đồng Tháp, [2] Trường Đại học Bách khoa Đà Nẵng, [3]Trường Đại học Cần Thơ

hoangntdt@gmail.com, hhhung@dut.udn.vn, hxhiep@ctu.edu.vn

TÓM TẮT: *Trong thời đại bùng nổ thông tin hiện nay, hệ tư vấn ngày càng có vai trò đặc biệt quan trọng và trở nên phổ biến trong việc hỗ trợ ra quyết định của con người. Trong các hệ tư vấn, các thuật toán lọc cộng tác là một trong những phương pháp phổ biến nhất để tạo khuyến nghị. Bên cạnh thành công của mình, các hệ tư vấn lọc cộng tác đều gặp vấn đề là chỉ đo ảnh hưởng của hai người dùng bằng các độ đo tương đồng dựa trên lịch sử tiêu dùng của họ và do đó, mặc định xem ảnh hưởng của cặp người dùng lên nhau là đối xứng, Tuy nhiên, trong khi thực tế, điều đó có thể không đúng, bởi lẽ người dùng có nhiều kinh nghiệm sẽ có mức độ ảnh hưởng lớn hơn người có ít kinh nghiệm hoặc những người mới bắt đầu tham gia hệ thống, tức là, sự ảnh hưởng qua lại giữa hai người dùng thường không đối xứng, chính sự khác biệt này có thể tạo ra các thiên lệnh nhất định trong các khuyến nghị của hệ tư vấn. Mặt khác, các hệ tư vấn truyền thống chủ yếu tập trung về tính chất logic thể hiện sự tồn tại hay không tồn tại mối quan hệ giữa người dùng và mục dữ liệu hay sản phẩm mà không quan tâm đến mối quan hệ ấy có thể chấp nhận đến mức độ nào. Để giải quyết các vấn đề trên, trong bài báo này, chúng tôi đề xuất một cách tiếp cận mới trong việc xây dựng hệ tư vấn theo kỹ thuật lọc cộng tác dựa trên việc phân tích biến thiên của chỉ số hàm ý trong trường hàm ý để xếp hạng và lọc thông tin dựa vào biến thiên của chỉ số hàm ý từ đó đưa ra một tư vấn có ý nghĩa, nhằm khắc phục nhược điểm của các hệ thống tư vấn truyền thống là không quan tâm đến mức độ ảnh hưởng bất đối xứng của người dùng và đưa ra khuyến nghị với một hàm ý nhất định để tạo ra một độ đo tương tự cho từng cặp người dùng.*

Từ khóa: *cường độ hàm ý thống kê, trường hàm ý, hệ tư vấn lọc cộng tác, độ đo bất đối xứng.*

I. GIỚI THIỆU

Ngày nay, các công cụ tìm kiếm đang đứng trước một thử thách ngày càng lớn: rất khó để tìm chọn thông tin hữu ích nhằm đưa ra quyết định với một số lượng lớn các lựa chọn trong một thời gian ngắn. Xu hướng chuyển dịch từ hoạt động tìm kiếm thông tin sang tư vấn khuyến nghị thông tin diễn ra nhanh chóng hơn bao giờ hết và hệ tư vấn [15] trở thành một công cụ cực kỳ cần thiết và được sử dụng rộng rãi nhiều trong thương mại, dịch vụ điện tử như Banes & Noble, Amazon, Pandora, Netflix, Walmart,... Mục tiêu của hệ tư vấn là để lọc các thông tin hữu ích từ một số lượng lớn các thông tin để có thể dự đoán được sự đánh giá mà một người dùng sẽ cung cấp cho một mục và từ đó khuyến nghị các mặt hàng phù hợp cho người dùng [15]. Trong số các thuật toán hệ tư vấn, thì thuật toán lọc cộng tác [15] là một trong những kỹ thuật thành công nhất. Tuy vậy, bên cạnh những thành công của các thuật toán lọc cộng tác, thì chúng cũng gặp phải một số vấn đề nội tại trong việc tạo ra các tư vấn là (1) xem mức độ ảnh hưởng của cặp người dùng lên nhau là đối xứng liên quan đến việc sử dụng các độ đo độ tương đồng đối xứng. Tuy nhiên, trong thực tế, vai trò và sự ảnh hưởng qua lại giữa hai người dùng thường không đối xứng. Điều này có thể tạo ra một thiên lệch nhất định trong kết quả tư vấn. Bởi vì người dùng có kinh nghiệm nhiều hơn (ví dụ như một chuyên gia vào hệ thống) sẽ có ảnh hưởng nhiều hơn lên người dùng còn lại, nhưng người còn lại (ít kinh nghiệm hơn, hoặc mới tham gia) sẽ không thể có mức độ ảnh hưởng tương ứng ngược lại. (2) Các hệ tư vấn hiện nay chỉ tập trung giải quyết về mặt logic mối quan hệ tồn tại hay không tồn tại mối quan hệ ưu tiên giữa người dùng và mục dữ liệu mà chưa quan tâm đến mức độ xuất hiện hay mối quan hệ hàm ý (implicative) trên cơ sở thống kê giữa người dùng và mục dữ liệu trong thực tiễn. Mối quan hệ hàm ý này còn được xem như là mối quan hệ mang tính chất tri thức giữa người dùng và mục dữ liệu cần ưu tiên tư vấn. Do vậy, sử dụng một tiếp cận về sự tương đồng bất đối xứng là một hướng thu hút nhiều quan tâm để giảm thiểu sự thiên lệch từ sự khác biệt nêu trên trong kết quả tư vấn.

Gần đây, đã có một số nghiên cứu xây dựng các giải pháp cho hệ tư vấn dựa trên độ tương đồng người dùng bất đối xứng, như tư vấn lọc cộng tác dùng kỹ thuật phân rã ma trận dựa trên các độ đo bất đối xứng [1] [2], tư vấn dựa trên ảnh hưởng người dùng bất đối xứng kết hợp với giá trị tầm quan trọng toàn cục của người dùng [16] nhằm giải quyết vấn đề ảnh hưởng bất đối xứng của người dùng trong hệ tư vấn. Một xu hướng mới nữa là ứng dụng lý thuyết phân tích hàm ý thống kê trong lĩnh vực hệ tư vấn, là nhằm giải quyết vấn đề ảnh hưởng người dùng bất đối xứng và giải quyết vấn đề đánh giá được mức độ xuất hiện hay mối quan hệ hàm ý (implicative) giữa người dùng và mục dữ liệu trong thực tiễn như mô hình hệ tư vấn dựa trên tiếp cận luật kết hợp và độ đo cường độ hàm ý (implication intensity) [13] 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 mối quan hệ ưu tiên giữa người dùng và mục dữ liệu hay sản phẩm. Trong mô hình này nhóm tác giả đặc biệt quan tâm đến tỷ lệ hay mối quan hệ hàm ý 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. Một nghiên cứu khác trong ứng dụng phân tích hàm ý thống kê vào hệ tư vấn là tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo gắn kết hàm ý thống kê [14] để tính độ tương tự cho từng cặp người dùng bất đối xứng trong phương pháp lọc cộng tác.

Trong bài báo này, chúng tôi tiếp tục sử dụng phân tích hàm ý thống kê [3] [4] [7] nhưng đề xuất một hướng tiếp cận mới dựa trên phân tích xu hướng biến thiên của chỉ số hàm ý trong trường hàm ý thống kê (SIFA- Statistical

Implication Field Analysis) để giải quyết vấn đề ảnh hưởng người dùng bất đối xứng và mối quan hệ hàm ý giữa người dùng và mục dữ liệu của các hệ tư vấn.

Bài viết được tổ chức thành 5 phần, phần đầu tiên là giới thiệu bối cảnh và các vấn đề cần giải quyết của hệ tư vấn hiện nay cũng như đề xuất hướng tiếp cận giải quyết, phần II trình bày các nội dung có liên quan đến phân tích hàm ý thống kê và các nghiên cứu mở rộng trong trường hàm ý, phần thứ III trình bày mô hình hệ tư vấn dựa vào sự biến thiên của chỉ số hàm ý trong trường hàm ý, phần kế tiếp là phần thực nghiệm mô hình với các kịch bản và cuối cùng là phần kết luận.

II. TRƯỜNG HÀM Ý THỐNG KÊ

A. Trường hàm ý thống kê

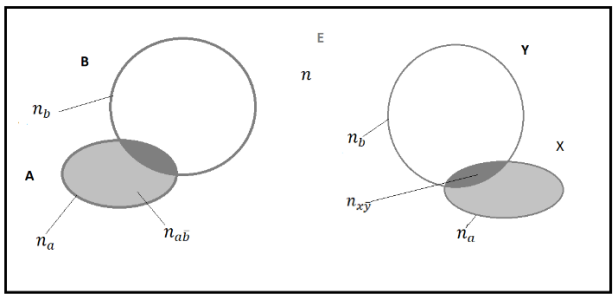
Phân tích hàm ý thống kê

Phân tích hàm ý thống kê (SIA- Statistical Implication Analysis) [3] [5] [7] [6], do Regis Gras đề xuất, nghiên cứu mối quan hệ hàm ý giữa các biến dữ liệu. SIA sử dụng độ đo cường độ hàm ý (hay còn gọi là chỉ số hàm ý Gras) để đo mối quan hệ hàm ý giữa các biến dữ liệu. Không giống như nhiều phương pháp phân tích dữ liệu khác, mối quan hệ giữa hai biến trong phương pháp này là không đối xứng. Độ đo trong phân tích hàm ý thống kê, cường độ hàm ý, 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, hoặc để đo sự tương quan giữa hai biến (cá nhân, thuộc tính…), độ đo này có tính chất là không đối xứng. Ban đầu, phân tích hàm ý thống kê đã được hình thành để đánh giá hành vi học tập của học sinh phổ thông trong quá trình giảng dạy toán học [3] [10], sau đó nó được ứng dụng trải rộng trên nhiều lĩnh vực như phân tích mối quan hệ tác động giữa các văn bản [7], trong bản thể học (ontology) [11], trong khoa học xã hội [9], tâm lý học [12].

Các luật không mang tính chất tuyệt đối mà mang tính chất hàm ý [3] [7] (có một xác suất nhất định) sẽ được phân tích dưới góc độ thống kê, gọi là phân tích hàm ý thống kê. Phân tích này cho phép phát hiện các luật (rules) không đối xứng theo dạng "nếu x thì gần như y" hay "nếu x thì y ở mức độ chấp nhận như thế nào". Các luật như vậy được gọi là luật hàm ý (implication rule) mà trong phạm vi bài viết này được gọi là luật để đơn giản trong trình bày.

Gọi E là tập hữu hạn các biến nhị phân, A và B là hai tập con của E, lần lượt là tập dữ liệu chứa các phần tử a $a \in A$ sao cho $A(a) = true$ và $b \in B$, sao cho B$(a) = true$, tập $\bar{A}, \bar{B}$ là tập bù của tập A và B tương ứng, gọi $n_a = card(A), n_b = card(B)$ là lực lượng của Tập A và B tương ứng, $n_{\bar{a}} = card(\bar{A}), n_{\bar{b}} = card(\bar{B})$ tương ứng là lực lượng của tập $\bar{A}$ và tập $\bar{B}$ và $n_{a\bar{b}} = card(A \cap \bar{B})$ là lực lượng của tập $A \cap \bar{B}$ là tập chứa các phần tử thỏa tính chất $a = true$ và $b = false$, $n_{a\bar{b}}$ còn gọi là số phản ví dụ (counterExample) hay bất khả dĩ (unlikelihood), đồng thời cũng chọn ngẫu nhiên và độc lập hai tập con X và Y có cùng lực lượng với A,B tương ứng, có nghĩa là $card(X) = n_a$ và $card(Y) = n_b$. Gọi $\bar{X}$và $\bar{Y}$ tương ứng là phần bù của X và Y trong E và có lực lượng tương ứng là $n_{\bar{a}} = n - n_a$ $n_{\bar{b}} = n - n_b$.

Mối quan hệ hàm ý giữa a và b được mô hình hóa trong phân tích hàm ý thống kê như sau (xem hình 1).



Hình 1. Minh họa các thành phần của phân tích hàm ý thống kê bằng biểu đồ venn

luật $a \rightarrow b$ is được chấp nhận ở ngưỡng α cho trước nếu:

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

Phân phối của $card(X \cap \bar{Y})$ phụ thuộc vào mẫu rút ngẫu nhiên [7], đối với một quá trình rút ngẫu nhiên cụ thể, mà trong đó các phần tử xuất hiện một cách linh động, theo dòng thực hiện của các giao dịch điền vào cơ sở dữ liệu. Quá trình này dừng lại khi có n_a phần tử với a = True, và n_b phần tử với b = True. Gán $card(X \cap \bar{Y})$ cho biến ngẫu nhiên số các phản ví dụ trong quá trình này. Có 3 giả thiết: (i) thời gian chờ đợi cho các sự kiện (a và $\bar{b}$) là các biến ngẫu nhiên độc lập , (ii) phân bố số lượng của các sự kiện đó xảy ra trong khoảng [t, t + T] chỉ phụ thuộc vào T , (iii) hai sự kiện có thể không đồng thời xảy ra. Khi đó, biến ngẫu nhiên $card(X \cap \bar{Y})$ theo phân phối Poisson $P(\lambda)$ với $\lambda = \frac{n_a n_{\bar{b}}}{n}$

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

Với $n_{\bar{b}} \neq 0$, gọi $Q(a, \bar{b})$ là biến ngẫu nhiên chuẩn tắc của $P(\lambda)$, thì $Q(a, \bar{b})$ được xác định bởi [3] [7]:

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

Trong thực nghiệm, giá trị quan sát của Q$(\boldsymbol{X}, \bar{\boldsymbol{Y}})$ là $\boldsymbol{q(X, \bar{Y})}$ được gọi là chỉ số hàm ý và xác định bởi:

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

Trong điều kiện xấp xỉ (ví dụ$\lambda \geq 4$), $Q(a, \bar{b})$ là xấp xỉ phân phối chuẩn $N(0,1)$.

Độ đo mật độ hàm ý $\varphi(a, b)$ của luật $a \rightarrow b$ xác định bởi **Error! Reference source not found.**: `

$$\varphi(a, b) = 1 - \Pr(\ Q(a, \bar{b}) \leq q(a, \bar{b})) = \begin{cases} 1 - \sum_{s=0}^{n_{a\bar{b}}} \frac{\lambda^s}{s!} e^{-\lambda} = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt \text{, với } n_y < n \\ 0, \ cho\ các\ trường\ hợp\ ngược\ lại \end{cases} \quad (4)$$

Luật hàm ý $X \rightarrow Y$ được chấp nhận ở mức tin cậy α nếu và chỉ nếu $\varphi(X, Y) \geq 1 - \alpha$. [7]

Cần chú ý là việc mô hình hóa mối quan hệ hàm ý này sẽ đo được sự "ngạc nhiên" trên cơ sở luật hàm ý có được. Do đó việc đánh giá mối quan hệ giữa các mục dữ liệu trong một hệ tư vấn dưới dạng luật hàm ý sẽ giúp xếp hạng hay đánh giá mối quan hệ cần có hay cần tư vấn giữa các mục dữ liệu.

Định nghĩa cường độ hàm ý này, gợi lại cho những người dùng độ hấp dẫn về hàm ý với điều kiện trên 0,50, có nghĩa là $q(a, \bar{b})$ phải âm. Vậy cũng có nghĩa là hàm càng âm thì có ý nghĩa hàm ý càng cao.

Với độ đo này thì $\varphi(a, b) \neq \varphi(b, a)$: phù hợp với quan điểm vai trò của người dùng là không đối xứng trong cùng một hệ thống.

Sự biến thiên của chỉ số hàm ý

Nghiên cứu sự ổn định của chỉ số hàm ý q [4], phải xem xét đến những biến đổi nhỏ ở vùng lân cận của toàn bộ 4 giá trị quan sát $(n, n_a, n_b, n_{a\bar{b}})$, để làm được điều này, cần phải thực hiện những mô phỏng khác nhau của toàn bộ 4 biến này, mà nó phụ thuộc vào q [4] [7]. Cần phải xem những biến này như những số thực và q giống như 1 hàm vi phân đối với các biến ràng buột theo bất đẳng thức: $0 \leq n_a \leq n_b$; $n_{a\wedge\bar{b}} \leq \inf\{n_a, n_b\}$ và $\sup\{n_a, n_b\} \leq n$. Thế thì sự khảo sát sự khác nhau của q đối với nhũng biến này và để khắc phục những hạn chế của toàn bộ giá trị của những tham số theo quan hệ $a \rightarrow b$. Sự khác biệt của q theo nghĩa hình học của Frechet được diễn đạt theo cách sau:

$$\mathrm{dq} = \frac{\partial \mathrm{q}}{\partial \mathrm{n}} \mathrm{dn} + \frac{\partial \mathrm{q}}{\partial \mathrm{n_a}} \mathrm{dn_a} + \frac{\partial \mathrm{q}}{\partial \mathrm{n_b}} \mathrm{dn_b} + \frac{\partial \mathrm{q}}{\partial \mathrm{n_{a\wedge\bar{b}}}} \mathrm{dn_{a\wedge\bar{b}}} = \mathrm{gradq.\,dM} \quad (5)$$

Với M là điểm có tọa độ $(n, n_a, n_b, n_{a\bar{b}})$ của trường vector vô hướng C, dM là vector thành phần vi phân của các biến thể hiện và grad q là vector đạo hàm riêng của các biến.

Vậy vi phân của hàm q xuất hiện giống như một tích vô hướng giữa gradient của nó và lượng tăng của q trên mặt biểu diễn các biến của hàm $q(n, n_a, n_b, n_{a\bar{b}})$. Cũng như gradient q biểu thị biến thiên bằng hàm hợp thành (từ 4 biến), là 4 lực lượng của các tập E, A, B và A ∩ $\bar{\mathrm{B}}$, chỉ ra hướng tăng của hàm q trong không gian 4 chiều. Nếu muốn nghiên cứu sự biến đổi của hàm q theo $n_{\bar{b}}$ cần phải thay thế $n_{\bar{b}}$ bởi $n - n_b$, và cần phải thay thế ký hiệu đạo hàm $n_{\bar{b}}$ trong đạo hàm riêng. Mặt khác, độ hấp dẫn của các vị trí khác nhau của nó trong việc đánh giá sự gia tăng của q (+ hoặc -) mà chúng ta ghi nhận là Δq đối với các biến thiên tương ứng $\Delta n, \Delta n_a, \Delta n_b, \Delta n_{a\bar{b}}$. Ta có:

$$\Delta q = \frac{\partial q}{\partial n} \Delta n + \frac{\partial q}{\partial n_a} \Delta n_a + \frac{\partial q}{\partial n_b} \Delta n_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}} \Delta n_{n_a\wedge\bar{b}} + o(\Delta q)$$

với o(Δq) là một vô cùng nhỏ

Xét đạo hàm riêng theo n_b và số phản ví dụ $n_{a\wedge\bar{b}}$. Ta được:

$$\frac{\partial q}{\partial n_b} = \frac{1}{2} n_{a\bar{b}} \left(\frac{n_a}{n}\right)^{\frac{1}{2}} (n - n_b)^{\frac{3}{2}} + \frac{1}{2} (\frac{n_a}{n})^{\frac{1}{2}} (n - n_b)^{\frac{1}{2}} > 0 \quad (6)$$

và:

$$\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_{a(n-n_b)}}{n}}} > 0 \quad (7)$$

Từ (6) và (7), chỉ số hàm ý đồng biến với các thành phần $n_a, n_{a\bar{b}}$, nếu các thành phần này tăng thì chỉ số hàm ý cũng tăng theo, do vậy cường độ hàm ý thì giảm.

$$\frac{\partial q}{\partial n_a} = -\frac{1}{2}\frac{n_{a\bar{b}}}{\sqrt{\frac{n_{\bar{b}}}{n}}}\left(\frac{n}{n_a}\right)^{\frac{3}{2}} - \frac{1}{2}\sqrt{\frac{n_{\bar{b}}}{n_a}} < 0 \quad (8)$$

Vậy những biến thiên của n_a trên $[0, n_b]$, thì chỉ số hàm ý q(a, b) thì luôn giảm dần (lỏm) đối với biến n_a và tối thiểu để $n_a = n_b$

Tiếp theo, mật độ hàm ý tăng và tối đa để na = nb, cho đạo hàm riêng q đối với n:

$$\frac{\partial q}{\partial n} = \frac{1}{2\sqrt{n}}\left(n_{a\wedge\bar{b}} + \frac{n_a n_{\bar{b}}}{n}\right) \quad (9)$$

Kết quả, nếu 3 tham số khác không thay đổi, chỉ số hàm ý giảm theo $\sqrt{n}$. Tính chất của hàm ý vì vậy là tốt hơn, đặc tính riêng biệt của A.S.I đối với những chỉ số khác cũng được xác nhận trong lĩnh vực chuyên môn [4]. Tính chất này được tán thành trong thống kê với độ tin cậy từ tần xuất quan sát.

Bây giờ, để khảo sát tiếp mối quan hệ giữa chỉ số hàm ý và cường độ hàm ý, hãy xem mật độ hàm ý $\varphi(a, b)$ như một hàm đối với $q(\mathrm{a}, \bar{b})$:

$$\varphi(\mathrm{q}) = \frac{1}{\sqrt{2\pi}}\int_{\mathrm{q}}^{\infty} \mathrm{e}^{-\mathrm{t}^2/2}\mathrm{dt}$$

Vậy thì, ta khảo sát sự thay đổi giá trị của $\varphi(\mathrm{q})$ khi giá trị q thay đổi trong vùng giá trị lân cận (a, b), giá trị q tự nó thay đổi như thế nào theo hàm 4 tham số xác định nó. Lấy nguyên hàm (lấy đạo hàm của của hàm tích phân 4 biến), ta có

$$\frac{\mathrm{d}\varphi}{\mathrm{dq}} = -\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\frac{-\mathrm{q}^2}{2}} < 0 \quad (10)$$

Điều này khẳng định mật độ hàm ý tăng khi q giảm, nhưng tốc độ tăng được xác định bởi công thức (10), điều này cho phép nghiên cứu kỹ hơn tính đúng đắn sự biến thiên của φ.

Trường hàm ý

a. Trường hàm ý thống kê

Chúng ta đến với chỉ số hàm ý $q(a, \bar{b})$. Xét không gian 4 chiều E, với các điểm M có tọa độ là những tham số liên quan đến các biến nhị phân a và b là $(n, n_a, n_b, n_{a\wedge\bar{b}})$, thì $q(a, \bar{b})$ là một trường vô hướng bằng việc áp dụng ánh xạ từ R^4vào R. Đối với vector grad q bao gồm những đạo hàm riêng của q đối với các biến $n, n_a, n_b, n_{a\wedge\bar{b}}$ là 1 trường gradient- trường vector đặc biệt mà chúng ta cũng gọi là trường hàm ý – vì nó thỏa mãn các tiêu chí Schwartz về vi phân hỗn hợp, nghĩa là phải thỏa mãn điều kiện đạo hàm hỗn hợp của từng cặp biến [4], tức là

$$\frac{\delta}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{b}}}\right) = \frac{\delta}{\delta \mathrm{n}_{\mathrm{b}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\right)$$

Từ (6) và (7), ta có:

$$\frac{\delta}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{b}}}\right) = \frac{1}{2}\left(\frac{\mathrm{n_a}}{\mathrm{n}}\right)^{-1/2}\left(\frac{\mathrm{n}_{\bar{\mathrm{b}}}}{\mathrm{n}}\right)^{-3/2} = \frac{\delta}{\delta \mathrm{n}_{\mathrm{b}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\right)$$

Tương tự như trên cho từng cặp biến khác trong bộ $(n, n_a, n_b, n_{a\wedge\bar{b}})$, ta cũng được kết quả:

$$\frac{\delta}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}}}\right) = \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\right); \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}}\right) = \frac{\delta}{\delta \mathrm{n}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\right); \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{b}}}\right) = \frac{\delta}{\delta \mathrm{n}_{\mathrm{b}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}}}\right); \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\right) = \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}\wedge\bar{\mathrm{b}}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}}}\right); \frac{\delta}{\delta \mathrm{n}_{\mathrm{a}}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}}\right) = \frac{\delta}{\delta \mathrm{n}}\left(\frac{\delta \mathrm{q}}{\delta \mathrm{n}_{\mathrm{a}}}\right).$$

Vậy, trường vector $C = (n, n_a, n_b, n_{a\wedge\bar{b}})$ trong E, phù hợp với trường gradient G và do đó nó là một trường hàm ý, nó được xem như là tiềm năng (potential) của q. Grad q là vector biểu diễn cho thay đổi không gian của mật của độ trường, nó sắp xếp các giá trị trường thấp đến giá trị cao hơn. Ở mỗi điểm của gradient, ta theo dõi sự gia tăng mật hàm ý của trong không gian và trong chừng mực nào tốc độ mà nó thay đổi dưới tác động của sự biến đổi của một hoặc nhiều tham số.

Ví dụ, nếu ta cố định 3 trong những tham số $n, n_a, n_b, n_{a\bar{b}}$ đã cho bởi việc thực hiện của cặp biến (a, b), gradient là 1 vector chỉ hướng tăng hoặc giảm của q, vậy việc giảm hoặc tăng của |q| và tiếp theo của hàm φ theo biến thiên của tham số thứ 4. Chúng ta chỉ ra thêm thông qua giải thích công thức (8).

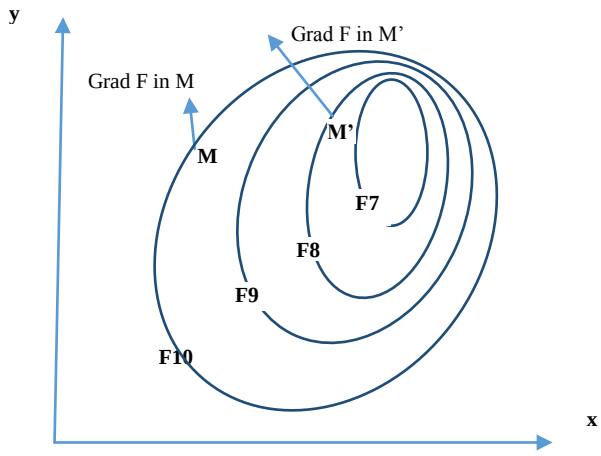
b. Mặt đẳng trị chỉ số hàm ý

Xét chỉ số hàm ý dưới dạng một hàm bốn biến $q(n, n_a, n_b, n_{a\wedge\bar{b}})$. Một đường hoặc mặt phẳng đẳng trị trong trường C thì cong trong E, không gian 4 chiều có thứ tự, dọc theo đó hoặc trên đó một điểm biến M duy trì cùng một giá trị của tiềm năng q. Phương trình của đường cong này là, được chỉ ra bởi **Error! Reference source not found.**:

$$q(a,\bar{b}) - \frac{n_{a\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = 0 \quad (11)$$

Tiếp theo, trên đường cong đó, tích vô hướng grad q.dM bằng 0 (theo (10) và (11)). Điều này được hiểu như là tính trực giao của tiếp tuyến gradient hoặc một tiếp tuyến siêu phẳng với đường cong đó, có nghĩa là với mặt phẳng đẳng trị.

Để dễ minh họa, ta xét mối quan hệ từ một potential F chỉ phụ thuộc vào 2 biến, hình dưới đây thông qua ví dụ cho thấy hướng trực giao của gradient đối với sự khác nhau của các mặt đẳng trị, dọc theo potentials F không đổi nhưng chuyển từ F7 đến F10 [4], như minh họa ở hình 2.



Hình 2. Minh họa mối quan hệ một potentials F chỉ phụ thuộc vào 2 biến

Trong trường hợp này, potential q lập nên những mặt đẳng trị như minh họa ở hình 2 (minh họa theo hai chiều để dễ biểu diễn). Ta có thể hiểu rằng trường này mạnh hơn đối với mặt phẳng chặt chẽ và yếu dần ở vùng thưa thớt hơn. Để có được 1 giá trị q trong trường hợp này, người ta cố định 3 biến, ví dụ như n, n_a, n_b.

Vậy, không gian E bao gồm nhiều lớp bởi các mặt đẳng trị tương ứng với những giá trị kế tiếp liên tục của q liên quan đến các lực lượng của , $n_a, n_b, n_{a\wedge\bar{b}}$. Tình huống này được đánh giá phù hợp trong mô hình của ASI [4].

Từ các kết quả trên, chúng tôi nhận thấy một số tính chất trong trường hàm ý:

- Trường hàm ý là một tập các mặt đẳng trị sinh ra do sự biến thiên của một biến bất kỳ trong bộ bốn các biến $(n, n_a, n_b, n_{a\wedge\bar{b}})$.

- Trên một mặt đẳng trị thì cường độ hàm ý là như nhau.

- Các mặt đẳng trị thì có thứ tự (do tính chất của trường gradient).

- Trường hàm ý có cường độ phụ thuộc vào mật độ của các mặt đẳng trị, tức là cường độ cao ở những nơi mà mật độ mặt đẳng trị chặc chẻ và giảm dần ở những nơi mặt đẳng trị thưa thớt hơn.

Cần chú ý là việc mô hình hóa mối quan hệ hàm ý này sẽ đo được sự "ngạc nhiên" trên cơ sở luật hàm ý có được. Do đó việc đánh giá mối quan hệ giữa các mục dữ liệu trong một hệ tư vấn dưới dạng luật hàm ý sẽ giúp xếp hạng hay đánh giá mối quan hệ cần có hay cần tư vấn giữa các mục dữ liệu.

III. MÔ HÌNH TƯ VẤN DỰA TRÊN ĐỘ BIẾN THIÊN CỦA CHỈ SỐ HÀM Ý TRONG TRƯỜNG HÀM Ý THỐNG KÊ

1. Luật hàm ý thống kê

Cho tập dữ liệu D là cơ sở dữ liệu bao gồm các giao dịch T, mỗi giao dịch Ti bao gồm các đối tượng hay mục dữ liệu(item) là đối tượng xuất hiện trong giao dịch như (sản phẩm, dịch vụ, hàng hóa,...). Tập mục dữ liệu (itemset) I là một tập bao gồm m mục dữ liệu. Tần số xuất hiện của một tập mục dữ liệu trong cơ sở dữ liệu được ký hiệu là δ. Độ hỗ trợ của tập X, ký hiệu là $support(x)$ là tỷ lệ giao dịch có chứa X trong cơ sở dữ liệu D. Luật kết hợp (association rule) là luật có dạng $X \rightarrow Y$, trong đó $X, Y \subset I$ là các itemsets, X được gọi là tiền đề (premise), Y là kết quả hay hệ quả (consequence). Luật kết hợp thường được đánh giá bằng hai độ đo là độ hỗ trợ (support - S) và độ tin cậy (confidence - C). Độ hỗ trợ (Support) của luật $X \rightarrow Y$ ký hiệu là $\sup(X \rightarrow Y)$ là tỷ lệ các giao dịch bao gồm cả X và Y trên tổng số giao dịch. $sup(X \rightarrow Y) = \frac{\delta(X \cup Y)}{|T|}$ Độ tin cậy (confidence) của luật $X \rightarrow Y$, ký hiệu: $conf(X \rightarrow Y)$

là khả năng mà một giao dịch chứa X thì sẽ chứa Y được xác định bởi: $conf(X \rightarrow Y) = \frac{\delta(X \cup Y)}{\delta(X)} = \frac{sup(X \rightarrow Y)}{sup(x)} = P(Y|X)$.

Ví dụ một tập dữ liệu luật kết hợp về các giao dịch mua sách với các luật kết hợp có confidence lớn hơn 50% và support >200 như bảng 1.

Bảng 1. Ví dụ tập dữ liệu mẫu của luật kết hợp

Rule no	Confidence (A=>C) %	Support ($A \cup C$)	Antecedent (A)	Consequent (C)
1	100	227	ItalCook=>	CookBks
2	62.77	204	ArtBks,ChildBks =>	GeoBks
3	54.13	203	CookBks, DoltYBks =>	ArtBks
4	61.98	207	ArtBks,CookBks =>	GeoBks
5	53.77	207	CookBks, GeoBks =>	ArtBks
6	57.11	245	RefBks =>	ChildBks, CookBks
7	52.31	204	ChildBks, GeoBks =>	ArtBks
8	60.78	203	ArtBks, CookBks =>	DoltYBks

Trong bảng, dòng số 2 (luật số 2) chỉ ra rằng nếu mua một cuốn sách nghệ thuật và một cuốn sách thiếu nhi, thì với sự tin cậy (confidence) 62,77% sách địa lý cũng sẽ được mua, cột support chỉ ra rằng luật có sự hỗ trợ của 204 giao dịch, có nghĩa là 204 người đã mua một quyển sách nghệ thuật, sách thiếu nhi và sách địa lý, tương tự cho các luật khác ở các dòng khác của bảng.

Luật hàm ý trong phân tích hàm ý thống kê được trình bày như sau: Giả sử, $A \subset I$ là tập những mục dữ liệu được đánh giá bởi người dùng u_a; $\bar{A}$ là tập bù của A. Tập $B \subset I$ là tập những mục dữ liệu được đánh giá bởi người dùng u_b; $\bar{B}$ là tập bù của B; $n_a = card(A)$ là số mục dữ liệu được đánh giá bởi người dùng u_a, là số phần tử của tập A); $n_b = card(B)$ là số mục dữ liệu được đánh giá bởi người dùng u_b (số phần tử của tập B); $n_{a\bar{b}} = card(A \cap \bar{B})$ là số mục dữ liệu được đánh giá bởi người dùng u_a nhưng chưa được đánh giá bởi người dùng u_b. Trong phân tích hàm ý thông kê, trình bày luật hàm ý dạng $A \rightarrow B$ là một dạng mỡ rộng của luật kết hợp không chỉ là biểu diễn mối quan hệ giữa các sự kiện mà còn các mối quan hệ giữa sự kiện với luật hoặc giữa các luật với nhau (hay còn gọi là siêu luật, R-luật) ngoài các độ đo đánh giá thông thường, thì như đã nêu trong đoạn trên, độ đo đặt thù cho luật hàm ý là chỉ số hàm ý và cường độ hàm ý, nhờ đó mà các luật hàm ý thể hiện được mức độ hàm ý mà các luật kết hợp không thể hiện được, một luật hàm ý được thể hiện bởi bộ bốn các giá trị $(n, n_a, n_b, n_{a\bar{b}})$. Chúng được gọi là lực lượng của luật hàm ý. Nói cách khác, luật hàm ý biểu diễn mối quan hệ hàm ý giữa người dùng u_a và u_b là mối quan hệ giữa tập mục A được thích bởi người dùng u_a và tập mục B được thích bởi người dùng u_b với một ngưỡng hàm ý nhất định, ngoài được biểu diễn bằng một bộ 4 phần tử trên.

2. Tư vấn dựa trên biến thiên của độ đo hàm ý thống kê

Để đưa ra một định nghĩa chính thức của nhiệm vụ tư vấn, cần phải giới thiệu một số khái niệm về hệ tư vấn. Theo đó, tập hợp của người dùng (users) trong hệ thống sẽ được ký hiệu là U, và tập các mục (items) bằng I. Hơn nữa, tập các xếp hạng (rating) trong hệ thống được biểu thị bởi R, và tập các giá trị (Scores) có thể có cho một đánh giá là S(ví dụ, $S = [1, 5]$ hay $S = \{like, dislike\}$). Ngoài ra, giả thiết rằng không có nhiều hơn một đánh giá có thể được thực hiện bằng cách người dùng bất kỳ $u \in U$ cho một mục hàng cụ thể $i \in \mathrm{I}$ và ghi vào $r_{ui} \in R$ cho đánh giá này. Để xác định tập con của người dùng u đã đánh giá một mục i, các ký hiệu U_i được sử dụng. Tương tự như vậy, I_u đại diện cho các tập hợp con của các mục hàng đã được đánh giá bởi một người dùng u. Cuối cùng, các mục hàng đã được đánh giá bởi hai người u, nghĩa là $\mathrm{I}_u \bigcap \mathrm{I}_v$, là một khái niệm quan trọng trong bài trình bày của bài viết, và I_{uv} được sử dụng để biểu thị khái niệm này, nghĩa là $\mathrm{I}_{uv} = \mathrm{I}_u \bigcap \mathrm{I}_v$. Trong một biểu diễn tương tự, U_{ij} được sử dụng để biểu thị tập hợp người dùng đã đánh giá cả các mặt hàng i và j, nghĩa là $U_{ij} = \mathrm{U}_i \bigcap \mathrm{U}_j$.

Hai trong số những vấn đề quan trọng nhất liên quan đến hệ thống tư vấn là mục hàng tốt nhất và đề xuất danh sách n mục dữ liệu tốt nhất cho người dùng. Vấn đề đầu tiên bao gồm trong việc tìm kiếm, đối với một người dùng cụ thể u, các mặt hàng mới $i \in \mathrm{I} \backslash \mathrm{I}_u$ mà người dùng u nhiều khả năng có quan tâm đến. Khi xếp hạng có sẵn, nhiệm vụ này thường được định nghĩa như là một hồi quy hay vấn đề phân loại (đa lớp) mà mục đích là để tìm hiểu một hàm

$$f: U \times I \rightarrow S$$

Mà dự đoán đánh giá $f(u, i)$ của một người dùng u cho một mục hàng mới i. Hàm này sau đó được sử dụng để giới thiệu cho người dùng tích cực u_a một một mục i^* mà đánh giá ước tính có giá trị cao nhất.

$$i^* = arg \max_{j \in I \backslash I_u} f(u_a, j) \quad (12)$$

Dựa vào các kết quả nghiên cứu về trường hàm ý nêu trên, chúng tôi đề xuất một thuật toán khuyến nghị như sau:

- *Thuật toán IRC (implication Rule Creation)*

Thuật toán này có tham số đầu vào là tập dữ liệu *dataSet*, là tập dữ liệu được truyền từ thuật toán RBEP sẽ được mô tả ở đoạn sau, kết quả là tập luật và bản số của tập luật dưới dạng bộ 4: $(n, n_a, n_b, n_{a\bar{b}})$.

Để sinh tập luật từ tập giao dịch dataSet, tập dữ liệu này được nhị phân hóa các đánh giá (rating) mục trên từng giao dịch của người dùng thành tập *data*. Để sinh luật trên tập *data*, có nhiều thuật toán khai khoáng dữ liệu cho việc sinh luật phổ biến hiện nay như thuật toán *apriori*,, *FP-Growth, setM, AIS (*Artificial immune system*)*..., trong đó thuật toán apriori dù không phải là thuật toán tốt nhất nhưng là một trong những thuật toán phổ biến và được quan tâm nhiều với nhiều phiên bản khác nhau như *aprioriTID*, *apriori Hybrid*. Trong bài viết này, thuật toán apriori được dùng[1] (do tính phổ dụng của nó và tập dữ liệu trong các kịch bản thực nghiệm trong bài viết này là không quá lớn để xử lý) với độ hỗ trợ và độ tin cậy tối thiểu là *minConf* và *minSup*. Trên mỗi luật được sinh ra, lần lượt xác định tập các bản số của chúng $(n, n_a, n_b, n_{a\bar{b}})$, n được xác định qua việc đếm số các giao dịch, các tham số còn lại được xác định bằng cách phân tích luật thành vế trái, vế phải sau đó áp dụng các phép toán ma trận đơn giản lên chúng, như mô tả trong thuật toán IRC.

Kết quả trả về là tập luật và bản số của chúng dùng cho thuật toán RBEP để sinh ra các luật hàm ý cùng với độ đo biến thiên chỉ số hàm ý trong trường hàm ý phục vụ cho mô hình tư vấn, và sẽ được trình bày ở đoạn sau.

Thuật toán IRC (Implication Rule Creation)

Input: tập các giao dịch (dataSet)
Output: tập luật và bản số của tập luật: $(n, n_a, n_b, n_{a\bar{b}})$

Begin
Chuyển Dataset dạng tập các mẫu tin người dùng nhị phân *data*
rules = apriori(data, minConf, minSup) #Dùng thuật toán apriori để tạo ra tập luật từ dataset.
Với mỗi luật i
{
Đếm số giao dịch (transaction) $n(i)$.
Tách tập luật thành hai ma trận vế trái $lhsRules$ và vế phải $rhsRules$

$$lhsRules[i,j] = \begin{cases} True, nếu\ item\ j\ thuộc\ về\ vế\ trái\ của\ luật\ i \\ False, trong\ trường\ hợp\ ngược\ lại \end{cases}$$

$$rhsRules[i,j] = \begin{cases} True, nếu\ item\ j\ thuộc\ về\ vế\ phải\ của\ luật\ i \\ False, trong\ trường\ hợp\ ngược\ lại \end{cases}$$

Tính $n_a(i)$:
$lhsProduct = data \times (lhsRules)^T$ # $(lhsRules)^T$ *ma trận chuyển vị của* $lhsRules$
$na[i] = rowSum(lhsProduct[i])$ # Đếm n_a của luật i
Tính n_b:
$rhsProduct = datar \times (hsRules)^T$
$nb[i] = rowSum(rhsProduct[i])$ # Đếm n_b của luật i
Tính n_{ab}: việc tính toán giống như cho $n_a$nhưng trên cả hai vế trái và vế phải của luật i
Tính $n_{a\bar{b}}$: $n_{a\bar{b}}(i) = n_a(i) - n_{ab}(i)$ # Đếm $n_{a\bar{b}}$ của mỗi luật i
}
Return cardinalities = $(n, n_a, n_b, n_{a\bar{b}})$
End

- *Thuật toán RBEP – Recommendation by equipotential plane*

Trong thuật toán này, các tham số đầu vào sẽ là tập dữ liệu *dataSet*, yếu tố biến thiên của chỉ số hàm ý *byFactor*, và ngưỡng θ cho mặt đẳng trị.

Tập dữ liệu *dataSet* sẽ được xử lý qua thuật toán IRC đã mô tả ở trên để sinh tập luật và các thông số (n, n_a, n_b, $n_{a\bar{b}}$) tương ứng với từng luật của nó, kế tiếp, trên mỗi luật chỉ số hàm ý $q(a,b)$ và đạo hàm riêng của nó theo yếu tố *byFactor* sẽ được tính toán. Không gian các giá trị đạo hàm riêng của chỉ số hàm ý sẽ được phân hoạch thành tập các mặt đẳng trị *recSet* với ngưỡng ($\left|\Delta q(a,\bar{b})\right| \leq \theta$).

Cuối cùng, dựa vào tập *recSet* để trả về kết quả tư vấn là *một mục* hay danh sách *k mục* phù hợp cho người dùng, như trình bày trong thuật toán RBEP dưới đây:

[1] Thuật toán Apriori có độ phức tạp về thời gian là $O(k * t * n)$ với k là kích thước tập mục phổ biến, t là kích thước tập dữ liệu và n là số tập mục của t (với $t >> k, n >> k$). Với độ phức tạp này, thuật toán phù hợp với các tập dữ liệu không quá lớn.

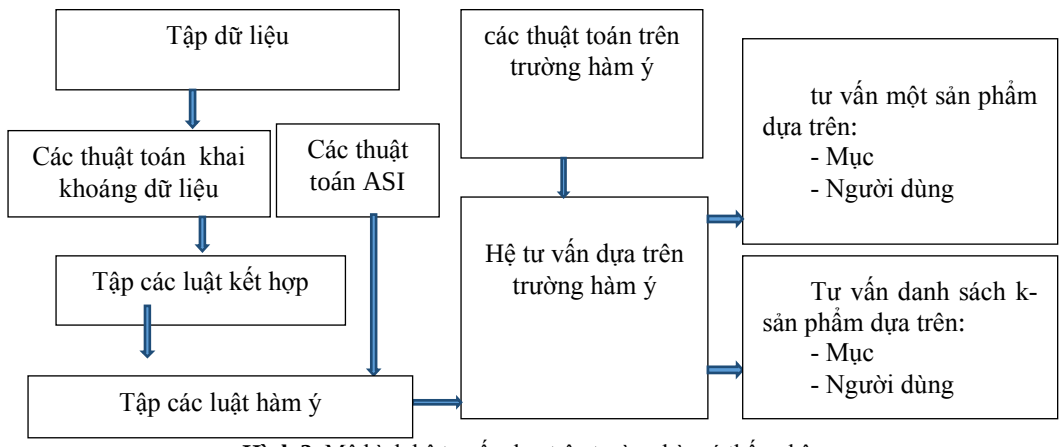
Thuật toán RBEP – Recommendation by equipotential plane

Input: dataSet, θ, byFactor ## byFactor is variant factor: $(n,\ n_a,\ n_b,\ n_{a\bar{b}})$
Output: recommendation item/ top k item list

Begin
Gọi hàm IRC(dataset) # sinh luật và tính $n, n_a, n_b, n_{a\bar{b}}$ của mỗi luật cho tập dataset bằng thuật toán IRC
Với mỗi luật rule(i),
{
Tính chỉ số hàm ý $q_i(a,b) = \frac{n_{a\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$
tính đạo hàm riêng $q_i(a,b)$ theo byFactor.
}
Xác định tập recSet chứa q trên cùng một mặt đẳng trị theo byFactor ($\left|\Delta q(a,\bar{b})\right| \leq \theta$)
Return mục or k mục từ tập recSet.
End

Thuật toán Rbep vừa mô tả trên làm cơ sở cho mô hình khuyến nghị cho hệ tư vấn dựa trên trường hàm ý thống kê (RSISF - Recommender System based on Implicative Statistical Field) như hình 3, theo đó, tập dữ liệu cần xử lý để đưa ra các tư vấn phù hợp cho người dùng sẽ được tiền xử lý để áp dụng các thuật toán khai khoáng dữ liệu để sinh ra các luật kết hợp, tiếp theo áp dụng các thuật toán phân tích hàm ý thống kê để có được tập luật hàm ý, cuối cùng dùng các thuật toán trên trường hàm ý RBEP và IRC để tạo ra trường hàm ý gồm nhiều mặt đẳng trị, từ đó tư vấn cho người dùng mục hay danh sách k-mục phù hợp với một mức độ hàm ý nhất định.

Mô hình này giúp người dùng giải quyết bài toán với mức độ hàm ý chấp nhận được ở một ngưỡng (tỷ lệ) α, nhằm dự báo cho người dùng 1 sản phẩm phù hợp nhất với yêu cầu hoặc liệt kê một danh sách k sản phẩm phù hợp nhất theo yêu cầu.



Hình 3. Mô hình hệ tư vấn dựa trên trường hàm ý thống kê

IV. THỰC NGHIỆM

1. Tập dữ liệu sử dụng

Với mô hình hệ tư vấn dựa trên dữ liệu phân tích hàm ý thống kê người dùng theo thời gian đã được đề xuất ở trên, chúng tôi tiến hành thực nghiệm trên bộ dữ liệu MovieLens đã được thu thập bằng cách nghiên cứu GroupLens từ trang web MovieLens[2] trong đó có khoảng 100.000 xếp hạng của khoảng 943 phim được thực hiện bởi 1.682 người dùng, các xếp hạng có giá trị từ 1-5 tương ứng với các bộ phim được đánh giá từ thấp đến cao nhất. Bộ dữ liệu được tiền xử lý nhằm phục vụ cho thực nghiệm được chính xác hơn, bằng cách:

- Chuẩn hóa dữ liệu: những người dùng xếp hạng cao (hoặc thấp) cho tất cả phim của họ tùy theo cá nhân có thể dẫn đến các thiên vị (bias) kết quả. Chúng ta có thể loại bỏ hiệu ứng này bằng cách chuẩn hóa dữ liệu sao cho đánh giá trung bình của mỗi người dùng là cùng một thang đo.

- Chọn dữ liệu có liên quan: bỏ qua các dữ liệu có thể dẫn đến các kết quả thiên lệch (bias) và cũng để tăng tốc độ tính toán, bằng cách không quan tâm đến các phim đã được xem chỉ một vài lần[3], vì giá trị xếp hạng của các bộ phim này có thể bị dẫn tới thiên lệch vì thiếu dữ liệu, và những người dùng đánh giá chỉ một vài bộ phim[4] vì xếp hạng của họ có thể là thành kiến.

[2] https://movielens.org/
[3] Chúng tôi chỉ quan tâm các bộ phim được đánh giá trên 90 lần
[4] Chúng tôi chỉ quan tâm đến những người dùng đã đánh giá trên 40 phim

Trên bộ dữ liệu đã được tiền xử lý như vậy và để tránh các vấn đề quá chuyên môn (overfitting), cũng như để có được độ chính xác tốt hơn chúng tôi tiến hành thực nghiệm theo phương thức k-fold cross validation, thay cho các phương pháp Splitting và Boostrapping là các phương pháp thực hiện việc thử nghiệm mô hình hệ tư vấn trên một phần của bộ dữ liệu (phần còn lại dùng làm tập kiểm tra), dù rằng khối lượng tính toán của phương pháp k-fold cross validation có thể tăng lên, do trong phương pháp này, bộ dữ liệu được chia thành k khối (chunk, fold), lấy một khối ra làm tập kiểm tra, và đánh giá độ chính xác, các khối còn lại dùng cho tập huấn luyện, sau đó lập lại quá trình này với tất cả các khối còn lại và tính toán độ chính xác trung bình của k lần kiểm tra, như vậy tập kiểm tra sẽ là toàn bộ dữ liệu chứ không trên một phần dữ liệu như hai phương pháp trước.

2. Công cụ thực nghiệm

Các thực nghiệm được thực hiện dựa trên gói công cụ *implicativefield* do chúng tôi xây dựng dựa trên ngôn ngữ R, bao gồm các công cụ nghiên cứu ứng dụng phân tích hàm ý thống kê phục vụ cho hệ tư vấn dựa trên độ biến thiên của chỉ số hàm ý trong trường hàm ý thống kê.

3. Kịch bản 1. Tư vấn dựa trên biến thiên chỉ số hàm ý theo một yếu tố trong trường hàm ý

Kết quả thực nghiệm của mô hình hệ tư vấn dựa trên độ biến thiên chỉ số hàm ý trong trường hàm ý trên tập dataset Movielens đã được tiền xử lý như nói ở đoạn trên, các tập luật được sinh ra (với điều kiện support = 0.4 và confident = 0.4), tổng cộng là 119 luật, sau khi loại bỏ các luật không có ý nghĩa (vế trái của luật bằng nil), và thỏa mãn cường độ hàm ý lớn hơn 0,5 ta chỉ chọn các luật có chỉ số hàm ý nhỏ hơn 0, còn lại 84 tập luật, Với ngưỡng $\theta = 0{,}5$, chúng tôi có được các giá trị cập nhật của chỉ số hàm ý q theo biến thiên của yếu tố bất kỳ của bộ 4 $(n, n_a, n_b, n_{a\wedge\bar{b}})$, trong kịch bản này byFactor= $n_{a\wedge\bar{b}}$ và thu được 18 tập các mặt đẳng trị trong trường hàm ý của mô hình là các siêu phẳng 3 chiều (n, n_a, n_b), các siêu phẳng này có mật độ phân bố tiềm năng (potentials) của chỉ số hàm ý là không điều nhau, được liệt kê theo bảng 3. trong đó mặt đẳng trị số 1 gồm các luật$(102, 119, 117, 95, 97, 70, 88)$, mặt đẳng trị số 2 (103), mặt đẳng trị số 3$(113, 98)$, mặt đẳng trị số 4 $(116, 118, 92, 101, 99)$,…, mặt đẳng trị số 18 $(85, 112, 82, 39)$. Các tập luật trên mỗi siêu phẳng có giá trị chỉ số hàm ý là như nhau với một ngưỡng xấp xỉ θ. Cụ thể siêu phẳng đẳng trị số 1 (7 luật, có chỉ số hàm ý cập nhật theo biến thiên = -8.94194$\pm\theta$, với ngưỡng xấp xỉ $\theta = 0{,}5$) như bảng 2.

Bảng 2. Mặt đẳng trị chỉ số hàm ý thứ 1 trên trường hàm ý sinh ra từ yếu tố byFactor = $n_{a\wedge\bar{b}}$

No	Mô tả luật	n	n_a	n_b	$n_{a\bar{b}}$	$\frac{\partial q}{\partial n_{a\bar{b}}}$	$q(a,b)$	$q(b,a)$
102	{Star Wars (1977),Empire Strikes Back, The (1980)} => {Raiders of the Lost Ark (1981)}	633	316	385	21	0.089873669	-9.239382357	-9.149508687
119	{Star Wars (1977),Raiders of the Lost Ark (1981),Return of the Jedi (1983)} => {Empire Strikes Back, The (1980)}	633	306	333	35	0.083038695	-9.136224171	-9.053185476
117	{Star Wars (1977),Empire Strikes Back, The (1980),Return of the Jedi (1983)} => {Raiders of the Lost Ark (1981)}	633	287	385	16	0.094305075	-9.095001955	-9.00069688
95	{Empire Strikes Back, The (1980),Return of the Jedi (1983)} => {Raiders of the Lost Ark (1981)}	633	290	385	17	0.093816022	-9.064287696	-8.970471673
97	{Raiders of the Lost Ark (1981),Return of the Jedi (1983)} => {Empire Strikes Back, The (1980)}	633	311	333	38	0.082368476	-9.010564805	-8.928196329
70	{Empire Strikes Back, The (1980)} => {Raiders of the Lost Ark (1981)}	633	333	385	29	0.087549546	-8.883166958	-8.795617412
88	{Return of the Jedi (1983)} => {Star Wars (1977)}	633	411	471	15	0.097504225	-8.7934024	-8.695898175

Có thể nhận thấy:

- Xu hướng biến thiên của hàm ý theo yếu tố byFactor, ở đây yếu tố $n_{a\wedge\bar{b}}$, có vai trò chính trong việc cũng cố hay từ chối một luật (theo lý thuyết hàm ý thống kê nêu ở đoạn trên), yếu tố này tăng lên đã làm tăng giá trị chỉ số hàm ý, đồng nghĩa với cường độ hàm ý giảm, tuy nhiên lượng giảm không đáng kể, do vậy tập luật trên mặt đẳng trị này vẫn còn ở mức hàm ý cũ. Điều này cũng nói lên được độ ổn định cao của chỉ số hàm ý thống kê, khi số lượng phản ví dụ tăng một lượng nhỏ thì luật vẫn còn ý nghĩa.

- Mật độ trường hàm ý phân bố không đều, mật độ hàm ý cao ở những mặt đẳng trị có giá trị chỉ số hàm ý biến thiên ít và tập trung được nhiều giá trị hơn như các mặt đẳng trị số 8, 15, 1 và 9. Mật độ trường hàm ý thưa dần và thấp nhất ở các mặt như mặt đẳng trị số, 2, 7, và 13 như trong bảng 3. Điều này, thể hiện sự phù hợp của luật với xu hướng biến thiên của chỉ số hàm ý, khi chỉ số hàm ý biến thiên đến một lượng nhất định mà khi đó luật không được chấp nhận ở một ngưỡng hàm ý xác định thì nó sẽ chuyển đến một mặt đẳng trị khác có ngưỡng hàm ý phù hợp hơn. Và do vậy, sẽ giúp cho việc tư vấn cho người dùng những mục dữ liệu có mức độ hàm ý phù hợp nhất.

Bảng 3. Mật độ của trường hàm ý trên các mặt đẳng trị và chỉ số hàm ý của nó

Mặt đẳng trị	Số luật	Chỉ số hàm ý	Mặt đẳng trị	Số luật	Chỉ số hàm ý
1	7	-8.94194	10	6	-4.9362
2	1	-8.69112	11	4	-4.47583
3	2	-8.1998	12	4	-4.07976
4	5	-7.75697	13	1	-3.83082
5	7	`-7.38222	14	3	-3.26428
6	5	-6.98373	15	9	-2.72022
7	1	-6.73528	16	4	-2.34934
8	9	-5.86112	17	4	-1.86901
9	7	-5.39429	18	4	-1.35299

- Một người dùng mục tiêu sẽ được khuyến nghị bộ phim hoặc danh sách các phim mà người ấy sẽ thích theo các nội dung luật tương ứng dựa vào các phim trước đây mà họ đã từng xem, như trong mặt đẳng trị số 1 ở bảng 1, người dùng đã xem {Star Wars (1977), Empire Strikes Back, The (1980)} thì gợi ý cho anh/chị ta là phim {Raiders of the Lost Ark (1981) với chỉ số hàm ý là -5.86122.

Trong kịch bản này, trường hàm ý cũng có thể được phân hoạch thành tập các mặt phẳng đẳng trị chứa các tập luật có cùng chỉ số hàm ý theo byFactor để làm cơ sở cho các kết quả tư vấn.

4. Kịch bản 2. Khuyến nghị dựa trên biến thiên chỉ số hàm ý theo đa yếu tố trong trường hàm ý

Tương tự như trong kịch bảng 1, kết quả được các giá trị cập nhật của chỉ số hàm ý q theo biến thiên đồng thời của byFactor = (n_b và $n_{a\wedge\bar{b}}$), và thu được 19 tập các mặt đẳng trị trong trường hàm ý của mô hình (là các mặt phẳng 2 chiều (n, n_a)), được liệt kê theo số hiệu luật như sau: $eqPn_{b_1} = (102, 119, 117, 95, 97, 70, 103)$, $eqPn_{b_2} = (88)$, $eqPn_{b_3} = (88, 113)$, $eqPn_{b_{18}} = (85, 112, 82, 83)$, $eqPn_{b\wedge\bar{b}_{19}} = (39)$. Cụ thể siêu phẳng đẳng trị số 1 (7 luật, có chỉ số hàm ý cập nhật theo biến thiên yếu tố hệ quả (n_b) là= -8.91526$\pm\theta$ với $\theta = 0{,}5$ như bảng 4:

Bảng 4. Trình bày mặt đẳng trị thứ 1 của chỉ số hàm ý theo yếu tố n_b và $n_{a\wedge\bar{b}}$ trên trường hàm ý

No	Mô tả luật	n	n_a	n_b	$n_{a\bar{b}}$	$\frac{\partial q}{\partial n_b}$	$\frac{\partial q}{\partial n_{a\wedge\bar{b}}}$	$q(a,b)$
102	{Star Wars (1977),Empire Strikes Back, The (1980)} => {Raiders of the Lost Ark (1981)}	633	316	385	21	0.026238057	0.089873669	-9.12327063
119	{Star Wars (1977),Raiders of the Lost Ark (1981),Return of the Jedi (1983)} => {Empire Strikes Back, The (1980)}	633	306	333	35	0.024914888	0.083038695	-9.028270588
117	{Star Wars (1977),Empire Strikes Back, The (1980),Return of the Jedi (1983)} => {Raiders of the Lost Ark (1981)}	633	287	385	16	0.024420896	0.094305075	-8.976275984
95	{Empire Strikes Back, The (1980),Return of the Jedi (1983)} => {Raiders of the Lost Ark (1981)}	633	290	385	17	0.024705711	0.093816022	-8.945765963
97	{Raiders of the Lost Ark (1981),Return of the Jedi (1983)} => {Empire Strikes Back, The (1980)}	633	311	333	38	0.025450948	0.082368476	-8.902745381
70	{Empire Strikes Back, The (1980)} => {Raiders of the Lost Ark (1981)}	633	333	385	29	0.028147259	0.087549546	-8.767470153
103	{Star Wars (1977),Raiders of the Lost Ark (1981)} => {Empire Strikes Back, The (1980)}	633	347	333	52	0.028131495	0.07797879	-8.662992663

Các luật nằm trên cùng siêu phẳng này sẽ làm cơ sở cho việc khuyến nghị một bộ phim hoặc danh sách tối đa 7 bộ phim cho người dùng mục tiêu phù hợp với chỉ số hàm ý -8.9126 có xét đến xu hướng biến thiên của nó theo yếu tố n_b và $n_{a\wedge\bar{b}}$. kết quả này có thể xem như việc phân hoạch trưởng hàm ý theo tố n_b và $n_{a\wedge\bar{b}}$ do vậy số mặt đẳng trị tăng lên do mức độ phân hoạch theo nhiều yếu tố hơn, điều này thấy rỏ hơn khi, phân hoạch theo 3 yếu tố, như trong thực nghiệm này chọn byFactor là $(n, n_b, n_{a\wedge\bar{b}})$, lúc này trường hàm ý được phân hoạch thành 35 mặt đẳng trị là các đường thẳng (mặt đẳng trị một chiều), các đường đẳng trị xuất hiện nhiều hơn, mật độ tiềm năng của chỉ số hàm ý thưa dần (xuất hiện nhiều tập luật với số lượng nhỏ và thậm chí chỉ 1 luật), như trong bảng 5.

Bảng 5. Tổng hợp kết quả biến thiên chỉ số hàm ý theo 3 yếu tố ($n, n_b, n_{a\wedge\bar{b}}$)

No	Số luật	No	Số luật	No	Số luật	No	Số luật
1	7	10	3	19	2	28	5
2	3	11	1	20	1	29	2
3	2	12	1	21	4	30	1
4	3	13	6	22	1	31	2
5	1	14	4	23	2	32	1

6	3	15	2	24	1	33	3
7	3	16	1	25	1	34	2
8	4	17	3	26	1	35	1
9	1	18	4	27	1		

Khi khảo sát biến thiên chỉ số hàm ý theo 4 yếu tố, lúc này, các siêu phẳng đẳng trị của chỉ số hàm ý là một điểm, số lượng siêu phẳng chính là số tập luật (cụ thể là 84), mỗi siêu phẳng chỉ chứa một luật, và giá trị biến thiên của chỉ số hàm ý đại diện cho xu hướng biến thiên hàm ý của chính luật đó.

Lúc này, tư vấn là những phim chỉ ra từ luật thoả mãn tất cả biến thiên của chỉ số hàm ý theo bốn yếu tố cấu thành và có cường độ hàm ý như mong đợi của người dùng.

Từ sự biến thiên mở rộng trên nhiều yếu tố (của bộ 4), Có thể thấy rằng:

- Trong trường hàm ý, các luật có một giá trị hàm ý (cường độ hàm ý) nhất định, khi có một biến thiên ở các yếu tố trong bộ bốn thì giá trị hàm ý của luật có thay đổi, nếu giá trị thay đổi vẫn chưa vượt qua ngưỡng hàm ý của mặt đẳng trị thì luật vẫn còn ý nghĩa trên mặt đẳng trị ấy, ngược lại luật sẽ bị chuyển sang một mặt đẳng trị khác có ngưỡng giá trị hàm ý phù hợp với giá trị hàm ý của luật và sẽ được dùng để tư vấn cho người dùng ở một bối cảnh có ý nghĩa khác.

- Mật độ trường hàm ý trong các mặt đẳng trị là biến thiên khi các yếu tố trong bộ 4 yếu tố biến thiên, điều này giúp cho việc tư vấn cho người dùng các mục theo ngữ cảnh thích hợp, giúp làm tăng độ chính xác và hiệu quả trong các khuyến nghị.

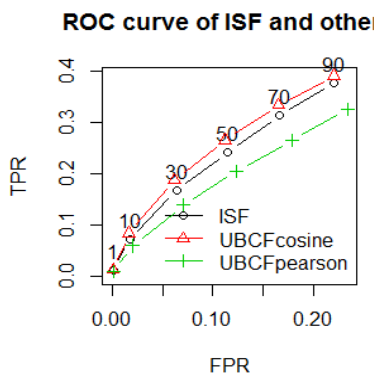
- Kết hợp với kết quả trong kịch bản 2, có thể thấy rằng trường hàm ý là một phân hoạch của các mặt đẳng trị sinh ra từ sự biến thiên của các yếu tố trong bộ 4.

5. Kịch bản 3. So sánh mô hình tư vấn dựa trên biến thiên chỉ số hàm ý trong trường hàm ý với các mô hình tư vấn lọc cộng tác dùng các độ đo tương đồng đối xứng

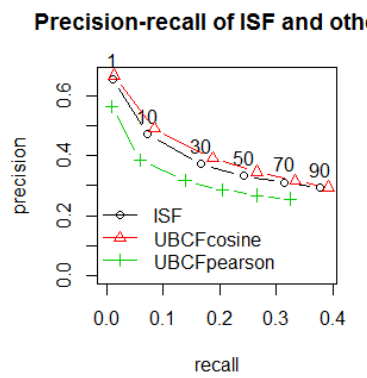
Trong kịch bản này, mô hình ISF được thực nghiệm cùng với các mô hình tư vấn lọc cộng tác dựa trên người dùng và trên mục dùng các độ đo tương đồng đối xứng, qua các chỉ số đánh giá lỗi khuyến nghị RMSE, MSE và MAE theo bảng 6, Mô hình ISF có các giá trị lỗi dự đoán là thấp hơn các mô hình khác. Về việc đánh giá độ chính xác với kết quả trong hình 4 và hình 5, độ chính xác của ISF là tốt tiệm cận (gần bằng) với mô hình UBCF dùng độ đo cosine và tốt hơn mô hình UBCF dùng độ đo pearson, nhưng với các mô hình IBCF (kết quả trong hình 6 và hình 7) thì độ chính xác của mô hình ISF là tốt hơn.

Bảng 6. Tổng hợp các chỉ số đánh giá lỗi của mô hình ISF và các mô hình IBCF và UBCF dùng độ đo đối xứng

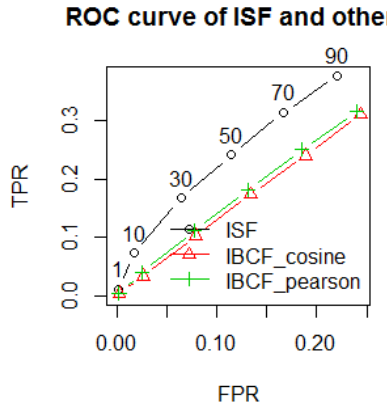
Mô hình	RMSE	MSE	MAE	Mô hình	RMSE	MSE	MAE
ISF	0.9488254	0.9002697	0.7425954	UBCF pearson	1.0047052	1.0094325	0.7910304
IBCF cosine	1.2441514	1.5479126	0.9296236	UBCF cosine	0.9918301	0.9837270	0.7775035
IBCF pearson	1.2204327	1.4894560	0.9162715				



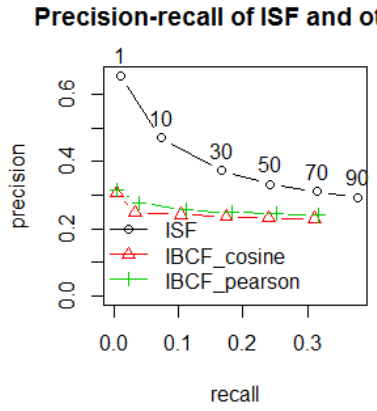
Hình 4. Đường cong ROC của mô hình ISF và các mô hình ISF dùng độ đo đối xứng



Hình 5. Precision và Recallcủa mô hình ISF và các mô hình ISF dùng độ đo đối xứng



Hình 6. Đường cong ROC của mô hình ISF và các mô hình IBCF dùng độ đo đối xứng



Hình 7. Precision và Recall của mô hình ISF và các mô hình IBCF dùng độ đo đối xứng

V. KẾT LUẬN

Lý thuyết phân tích hàm ý thống kê đã mở ra một hướng nghiên cứu về việc xây dựng mô hình hệ thống tư vấn dựa trên trường hàm ý. Trong mô hình này, chúng tôi xây dựng một độ đo mới có tính bất đối xứng và phản ánh độ tương đồng với một hàm ý nhất định, độ đo biến thiên của chỉ số hàm ý trong trường hàm ý, cho phép việc xếp hạng và lọc dữ liệu theo các mặt đẳng trị (mỗi mặt đẳng trị là một tập các luật hàm ý với cùng một đo chỉ số hàm ý xác định và các mặt này có thứ tự) để làm cơ sở cho hệ tư vấn. Kết quả nghiên cứu đã tạo được một mô hình tư vấn dựa trên trường hàm ý *implicativefield* cùng các thuật toán tư vấn lọc cộng tác dựa trên độ biến thiên chỉ số hàm ý trong trường hàm ý mang tính bất đối xứng, công cụ này đã được thực nghiệm trên tập dữ liệu MovieLens để xây dựng mô hình tư vấn và kiểm tra, so sánh với các mô hình tư vấn lọc cộng tác truyền thống dùng các độ đo tương đồng đối xứng, và cho kết quả tốt. Các đóng góp này nhằm mục đích làm tăng hiệu quả của các khuyến nghị (giúp tăng thêm tính ổn định và độ chính xác của dự báo xếp hạng và chỉ ra được xu hướng của các luật).

TÀI LIỆU THAM KHẢO

[1] Bin Cao, Qiang Yang, Jian-Tao Sun, Zheng Chen, Learning bidirectional asymmetric similarity for collaborative filtering via matrix factorization, Data Mining and Knowledge Discovery, Volume 22, Issue 3, pp.393–418, 2011.

[2] Rahul Katarya, Om Prakash Verma, Effectivecollaborative movie recommender system using asymmetric user similarity and matrix factorization, The 2016 IEEE International Conference on Computing, Communication and Automation (ICCCA'16), DOI: 10.1109/CCAA.2016.7813692, pp.1-12, 2016.

[3] Régis Gras et al., L'implication statistique - Nouvelle méthode exploratoire de données, La pensée sauvage édition, 1996.

[4] Régis Gras, Pascale Kuntz and Nicolas Greffard, Notion de champ implicatif en analysis statistique implicative, The 8th International Meeting on Statistical Implicative Analysis, Tunisia, pp 1-21, 2015 (in French).

[5] Régis Gras, Dominique Lahanier-Router, Duality between variables space and subjects space of the statistic implicative analysis, Dualite entre espace des variables et espace des sujets en analyse statisticque implicative, The VI International conference, ASI Analyse statistique implicative- Implicative statistical Analysis Caen (ASI6), France, pp 1-28, 2012.

[6] Régis Gras, Pascale Kuntz, Discovering R-rules with a directed hierarchy,Journal Soft Computing - A Fusion of Foundations, Methodologies and Applications (Volume 10 Issue 5), Springer-Verlag, pp 453-460, 2006.

[7] Régis Gras, Einoshin Suzuki Fabrice Guillet, Filippo Spagnolo (Eds.), Statistical Implicative Analysis, Theory and Application, Springer Verlag Berlin Heidelberg, 2008.

[8] Regis Gras, Pascale Kuntz. and Briand H., "Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données", The Mathématiques et Sciences Humaines 39, pp.9-29, 2001.

[9] Regis Gras, Raphael Couturier, Spécificités de l'Analyse Statistique Implicative (A. S. I.) par rapport à d'autres mesures de qualité de règles d'association, Quaderni di Ricerca in Didattica - GRIM (ISSN on-line 1592-4424), Eds : J. C. Régnier, R.Gras, F. Spagnolo, B. Di Paola, Université de Palerme, p.19-57, 2010. [in French].

[10] Dominique Lahanier-Reuter, Didactics of Mathematics and Implicative Statistical Analysis, Statistical Implicative Analysis - Studies in Computational Intelligence (Vol. 127), pp 277-298, 2008.

[11] Jérôme David, Fabrice Guillet, Régis Gras et Henri Briand, On the use of Implication Intensity for matching ontologies and textuel taxonomies, Statistical Implicative Analysis (R. Gras, E. Suzuki et F. Guillet éds.), Springer, pp.227-246, 2008.

[12] Jérôme David, Fabrice Guillet, Vincent Philippé, and Régis Gras, Implicative statistical analysis applied toclustering of terms taken from a psychologicaltext corpus, Proceedings of the 11th symposium on Applied Stochastic Models and Data Analysis (ASMDA 05), pp.201-208, 2005.

[13] Phan Quốc Nghĩa, Nguyễn Minh Kỳ, Nguyễn Tấn Hoàng, Huỳnh Xuân Hiệp, Hệ tư vấn dựa trên tiếp cận luật kết hợp và độ đo hàm ý thống kê, Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, 2015.

[14] Phan Phương Lan, Trần Uyên Trang, Huỳnh Hữu Hưng, Huỳnh Xuân Hiệp, Tư vấn lọc cộng tác dựa trên người sử dụng dùng phép đo gắn kết hàm ý thống kê, Kỷ yếu Hội nghị Quốc gia lần thứ IX về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'15); Cần Thơ, 2016.

[15] 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.

[16] Zhi-Lin Zhao Chang-Dong Wang, Jian-Huang Lai AUI&GIV Recommendation with asymmetric user influence and global importance value. Public Library of Science ONE, pp.2016.

RECOMMENDATION BASED ON THE VARIANCE OF IMPLICATION INDEX IN IMPLICATION FILED

Nguyen Tan Hoang, Huynh Huu Hung, Huynh Xuan Hiep

ABSTRACTS: *In the age of information explosion, the Recommender systems have become increasingly important and popular in supporting human decision-making. In the Recommender Systems, collaborative filtering algorithms are one of the most popular methods to create recommendations. In the collaborative filtering algorithms, a similarity measure plays a crucial role in making the recommendation, according to which, the popular measure be used today that are all symmetrical, and therefore the default, the effect of a pair of users is the same. However, in practice, that may not be true. The more experienced a user is, the more likely he or she is to have a greater influence than the less experienced or beginner, that is, the interaction between the two users is often asymmetric. This difference can lead to bias in recommendations of recommender systems. On the other hand, traditional recommender systems mainly focus on the logic of the existence or non-existence of a relationship between a user and an item without regard how to them being admissible (for a given threshold).To address these issues, we propose a new approach in developing collaborative filtering recommendation systems based on the implication of field analysis to rank and filter information based on implication index variations to limit the disadvantages of traditional recommendation systems, and to evaluate it against the recommended models using previous similarity measures.*