Một mô hình khám phá cộng đồng người dùng trên mạng xã hội

- **Hồ Trung Thành** [1]
- **Đỗ Phúc** [2]

[1] Khoa Hệ thống thông tin, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, thanhht@uel.edu.vn

[2] Trường Đại Học Công Nghệ Thông Tin, ĐHQG-HCM, phucdo@uit.edu.vn

(Bản nhận ngày 13 tháng 04 năm 2015, hoàn chỉnh sửa chữa ngày 08 tháng 04 năm 2016)

TÓM TẮT

Xu thế phát triển công nghệ và ngày càng xuất hiện nhiều loại hình truyền thông mạng xã hội dẫn đến sự thay đổi về hành vi của con người trong xã hội và hình thành những cộng đồng trực tuyến. Hành vi con người thay đổi dẫn đến nhiều hình thức kinh doanh, tiếp thị, dịch vụ và kể cả trong lĩnh vực giáo dục, an ninh, chính trị cũng thay đổi theo từ cách tiếp cận cho đến việc quản lý người dùng. Cộng đồng người dùng mạng xã hội ảnh hưởng và chi phối hành vi, thói quen của từng người dùng tham gia vào cộng đồng. Chính vì vậy, khám phá cộng đồng mạng xã hội từ nhiều nguồn dữ liệu khác nhau thông qua việc phân tích nội dung trao đổi sẽ biết được cộng đồng những người dùng có những hành vi được thể hiện trong nội dung và chủ đề mà người dùng quan tâm trao đổi trong những thông điệp. Trong bài báo này, chúng tôi đề xuất mô hình mới khám phá cộng đồng người dùng trên mạng xã hội dựa theo mô hình chủ đề kết hợp phương pháp mạng Kohonen.Trong đó mô hình đề xuất tập trung khám phá cộng đồng mạng xã hội và phân tích sự thay đổi chủ đề quan tâm của người dùng trong lĩnh vực giáo dục trên mạng xã hội theo từng giai đoạn thời gian.

Từ khóa: *chủ đề, mô hình chủ đề, khám phá cộng đồng, phân tích sự thay đổi, mạng Kohonen, TART.*

1. GIỚI THIỆU

Cộng đồng là một tập thể cùng sống và làm việc trong cùng một môi trường [2][8][15][23] [24]. Cộng đồng mạng xã hội là một tập hợp các cá nhân tương tác thông qua các phương tiện truyền thông cụ thể, có khả năng vượt qua những ranh giới địa lý và chính trị để theo đuổi lợi ích hay mục tiêu chung. Một trong những loại hình cộng đồng ảo phổ biến nhất là cộng đồng trên mạng xã hội.Trong phạm vi nghiên cứu này, chúng tôi đề cập đến cộng đồng người sử dụng trên mạng xã hội.

Hình 1. Cộng đồng trên mạng xã hội1

Có thể định nghĩa, cộng đồng là một nhóm người dùng trong mạng xã hội có sự tương tác

[1] http://treeintelligence.com/en/influence-and-viralization-networks/

nhau và thường quan tâm đến chủ đề được thảo luận trong nhóm hơn những nhóm khác [11][14] [23]. Trong nghiên cứu này, tập hợp các cộng đồng trên mạng được ký hiệu là C và một cộng đồng đang xét được ký hiệu là c, như vậy $c \in C$.

Xác suất điều kiện của một cộng đồng người dùng biểu thị cho mức độ tham gia, cùng quan tâm chủ đề của người dùng trong cộng đồng [23]. Cụ thể, *p(c|u)* là xác suất của cộng đồng *c* có chứa người dùng *u*[2] (xem công thức (1)). Như vậy, người dùng u có thể thuộc một hay nhiều cộng đồng.

$$\sum_{c \in C} P(c|u) = 1 \quad (1)$$

Chủ đề quan tâm của người dùng thường thay đổi, điều này dẫn đến cộng đồng mạng xã hội cũng thường thay đổi theo. Việc chi phối dẫn đến sự thay đổi trong cộng đồng mạng có 2 nguyên nhân chính: (1) là hình thành hay thay đổi từ nhóm các bạn bè biết trước và cùng kết bạn trên mạng hoặc thông qua sự giới thiệu bạn bè cùng kết bạn; (2) là thông qua sở thích của từng người dùng trên mạng cùng kết bạn với nhau hoặc cùng quan tâm đến những chủ đề dựa trên nội dung thông điệp mà người dùng quan tâm trao đổi. Như vậy, mối quan hệ của cộng đồng mạng thông qua sở thích được xem như một mạng lưới với sự liên kết những thành viên và mối quan hệ thể hiện trên mạng xã hội [1][2][3][9][10]. Bởi vì những thông tin nội dung chính là những thuộc tính của từng thành viên trên mạng xã hội. Những nội dung thông tin này được tồn tại dưới dạng văn bản, hình ảnh,...Cùng một cộng đồng mạng có thể quan tâm trao đổi nhiều chủ đề trong một giai đoạn thời gian và một chủ đề cũng có thể có nhiều cộng đồng quan tâm trao đổi. Nhiệm vụ nghiên cứu đặt ra là làm thế nào để có thể khám phá nhằm tìm ra cộng đồng mạng cùng quan tâm đến những chủ đề thông qua những nội dung thông điệp được trao đổi của tập người dùng trong cộng đồng và từng chủ đề cụ thể có những cộng đồng nào quan tâm trao đổi?

Một thách thức nữa đặt ra là cộng đồng mạng thường xuyên thay đổi các thành phần trong mạng theo thời gian, chẳng hạn như: sự thay đổi số thành viên trong công đồng, chủ đề mà cộng đồng quan tâm trao đổi,...Chính vì vậy, thành phần thay đổi trong cộng đồng mạng thường liên quan đến một hay nhiều chủ đề mà cộng đồng mạng quan tâm, số lượng thành viên tham gia cộng đồng, mức độ quan tâm đến từng chủ đề tại từng thời điểm, và đặc biệt hơn nữa là sự thay đổi trong cộng đồng mạng ảnh hưởng rất nhiều vào hành vi, sự quan tâm và trao đổi của thành viên trong cộng đồng. Điều này đã thu hút rất nhiều nhà nghiên cứu quan tâm nhằm phân tích và truy vết thông tin lan truyền để tìm ra nguồn gốc của thông tin của người đăng (gửi) [11][30] hay tìm ra sự ảnh hưởng của người hay chủ đề quan trọng để phục vụ cho những chiến lược phát triển như quản lý cộng đồng người dùng mạng xã hội của công ty, tổ chức hay của một quốc gia; hiểu người dùng để thực hiện chiến lược marketing hiệu quả, quảng bá ngành nghề và môi trường đào tạo lĩnh vực giáo dục,...

Để có thể khám phá cộng đồng người dùng theo chủ đề theo từng giai đoạn thời gian, trong nghiên cứu này chúng tôi tiếp cận theo mô hình chủ đề nhằm khai thác khả năng phân tích nội dung tìm ra từng chủ đề trong từng nội dung thông điệp cùng với tập từ đặc trưng cho chủ đề [4][5][10][27][28] và tiếp tục khai thác hiệu quả mô hình TART khám phá cộng đồng theo chủ đề quan tâm của người dùng có yếu tố thời gian được chúng tôi đề xuất và giới thiệu trong nghiên cứu [16].

Bên cạnh việc khai thác hiệu quả mô hình TART, trong nghiên cứu này chúng tôi đề xuất mô hình khám phá cộng đồng người dùng trên mạng xã hội bằng phương pháp huấn luyện mạng Kohonen [17][27] kết hợp với mô hình TART. Tiếp sau đó, chúng tôi tập trung phân tích sự thay

đổi chủ đề và thành viên của cộng đồng theo từng giai đoạn thời gian.

Các phần tiếp theo của bài báo: phần 2 trình bày các nghiên cứu liên quan, phần 3 trình bày mô hình đề xuất khám phá cộng đồng người dùng trên mạng xã hội vàkhảo sát sự thay đổi chủ đề quan tâm và người dùng của cộng đồng theo từng giai đoạn thời gian, phần 4 trình bày thử nghiệm và kết quả, phần 5 kết luận, hướng phát triển và cuối cùng là tài liệu tham khảo.

2. CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Mô hình Group-Topic (GT)

Mô hình GT [1] quan tâm đến phương pháp gom nhóm người dùng theo chủ đề dựa trên thuộc tính và nội dung trao đổi của từng thành viên trên mạng. Áp dụng mô hình chủ đề với yếu tố bổ sung là nhóm (group) với phương pháp học không giám sát, mô hình GT xem mỗi thành viên có mối quan hệ với thành viên khác trên mạng nếu những thành viên đó có cùng hành vi trong một sự kiện và sự liên kết các nội dung văn bản với nhau trong cùng sự kiện đó. Hơn thế nữa, mô hình GT cho rằng mỗi sự kiện tương ứng với một chủ đề *T*. Chính vì vậy, nhóm thành viên trên một cấu trúc mạng (hay nhóm thành viên) không tồn tại lâu mà sẽ thay đổi những chủ đề khác nhau trong những sự kiện khác nhau [1]. Nghiên cứu chi tiết của mô hình GT đã đề xuất phương pháp khám phá các nhóm thành viên trên mạng theo chủ đề tiếp cận theo phương pháp mạng Bayesian.

2.2 Mô hình Community-User-Topic (CUT)

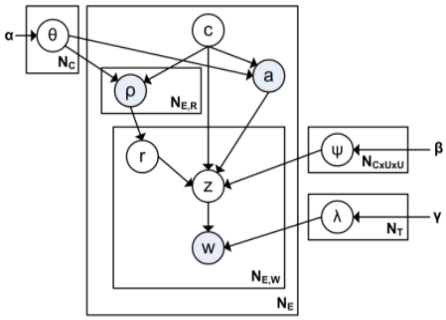
Trong nghiên cứu [3], nhóm tác giả giới thiệu mô hình CUT (C là công đồng – U là người dùng – T là chủ đề), trong đó tập trung nghiên cứu và đề xuất phương pháp khám phá cộng đồng dựa trên nội dung trao đổi và [3] cũng đã đề xuất hai mô hình thuộc CUT là CUT_1 và CUT_2. Mô hình CUT_1 và CUT_2 khác biệt nhau tại vị trí của tham số z và α_{di}. Kết hợp phương pháp mô hình xác suất và khám phá cộng đồng, nhóm tác giả trong [3] đề xuất ba thành phần chính là C, U, T trong mô hình.

Mục đích của hai mô hình này là rút trích cộng đồng người dùng theo chủ đề dựa trên dữ liệu, trao đổi trên mạng xã hội. Mô hình này dựa trên mạng Bayesian và Gibb sampling. Tuy nhiên, vì độ phức tạp của Gibb sampling, nên nhóm tác giả đã đề xuất ý tưởng đưa Gibb sampling kết hợp với việc lọc entropy để lưu vết quá trình thực hiện lấy mẫu và lọc từ, từ đó giúp cho quá trình thực hiện của mô hình sẽ không cần quan tâm đến những từ đã được xét đến hoặc những từ không có nghĩa.

2.3 Mô hình Community-Author-Recipient-Topic (CART)

Trong nghiên cứu [2], nhóm tác giả giới thiệu mô hình CART (Cộng đồng – Tác giả - Người nhận - Chủ đề), mô hình được thử nghiệm trên hệ thống dữ liệu Enron email. Mô hình chỉ ra rằng, sự thảo luận, trao đổi giữa những thành viên trong phạm vi một cộng đồng có liên quan đến những thành viên khác trong cùng cộng đồng. Mô hình này ràng buộc tất cả thành viên có liên quan và những chủ đề được thảo luận trong email thuộc về một cộng đồng, trong khi cùng những thành viên giống nhau và những chủ đề khác nhau có thể được gắn với cộng đồng khác. So sánh với các mô hình trên bao gồm cả CUT, mô hình CART lập luận chặt chẽ hơn để nhấn mạnh hơn nữa cách mà các chủ đề và mối quan hệ cùng ảnh hưởng đến cấu trúc của cộng đồng mạng trong vấn đề khám phá công đồng mạng theo chủ đề.

Mô hình CART [2] là một trong những cố gắng đầu tiên về nghiên cứu khám phá cộng đồng bằng sự kết hợp nghiên cứu dựa trên nội dung thông điệp mà thành viên trong cộng đồng mạng cùng trao đổi. Mô hình CART gồm 4 thành phần chính là C, A, R và T. Trong đó, C là cộng đồng người dùng, R là người nhận thông điệp, A là người gửi thông điệp, Z là chủ đề, W là từ thuộc chủ đề Z (hình 2) [2].



Hình 2. Mô hình CART [2]

Mô hình CART thực hiện theo các bước sau đây:

1. Sinh một dữ liệu email e_d, một cộng đồng c_d được chọn ngẫu nhiên
2. Dựa trên cộng đồng c_d, một người gửi a_d và tập người nhận ρ_dđược chọn
3. Sinh mỗi từ $w_{d,i}$trong dữ liệu email, một người nhận $r_{d,i}$ được chọn theo cách ngẫu nhiên từ tập người nhận ρ_d.
4. Dựa trên cộng đồng c_d, người gửi a_d và người nhận $r_{d,i}$ thì một chủ đề $z_{d,i}$ được chọn.
5. Từ $w_{d,i}$ được chọn dựa trên chủ đề $z_{d,i}$.

Kỹ thuật Gibb sampling cho mô hình CART như sau:

$$p(c_d, a_d, \rho_d, r_d, z_d, w_d)$$
$$= p(c_d)p(, a_d|c_d)\prod_{r\in\rho_d} p(r|c_d)\prod_{i=1}^{N_d} p(w_{d,i}|z_{d,i})$$
$$p(z_{d,i}|c_d, a_d, r_d, i) \qquad (2)$$

Trong đó, ρ_d là tập quan sát người nhận R, r_d là tập người nhận cần tìm (chọn từ ρ_d) and z_d là chủ đề tiềm ẩn thứ i tương ứng với mỗi từ thứ i $w_{d,i}$trong dữ liệu d, và N_d là tập từ trong dữ liệu.

2.4 Nhận định và động cơ nghiên cứu

Trong các nghiên cứu được giới thiệu, các nghiên cứu [1][2][3][13] trình bày trên và một số nghiên cứu khác như [6][7][24][25][26] đã đạt hiệu quả trong quá trình khám phá cộng đồng mạng dựa trên phân tích nội dung thông điệp. Tuy nhiên, các nghiên cứu này chưa quan tâm nhiều đến yếu tố thời gian cũng như chưa quan tâm đến việc phân tích sự thay đổi chủ đề quan tâm của người dùng thuộc cộng đồng theo thời gian. Bởi vì, sự thay đổi chủ đề quan tâm người dùng mạng có thể ảnh hưởng đến sự thay đổi chủ đề quan tâm của cộng đồng cũng như có thể thay đổi các thành phần trong cộng đồng mạng, chẳng hạn như khu vực địa lý hình thành cộng đồng, số thành viên tham gia, thời gian và chủ đề mà cộng đồng quan tâm trao đổi. Bên cạnh đó, vấn đề phân tích sự phân bố chủ đề trong cộng đồng mạng theo thời gian, phân bố chủ đề được quan tâm trong cộng đồng, với một chủ đề thì sự quan tâm của nhiều người dùng thay đổi ra sao, điều này cũng chưa được các nghiên cứu quan tâm. Hơn thế nữa, các nghiên cứu trên chủ yếu tập trung khám phá cộng đồng mạng trên tập ngữ liệu văn bản tiếng Anh, việc khai thác trên tập ngữ liệu văn bản tiếng Viết có nhiều khó khăn đặc biệt là hệ thống Tree Bank tiếng Việt còn chưa bao quát hết hệ thống từ trong tiếng Việt, từ ghép, từ đa nghĩa,...

3. MÔ HÌNH KHÁM PHÁ CỘNG ĐỒNG

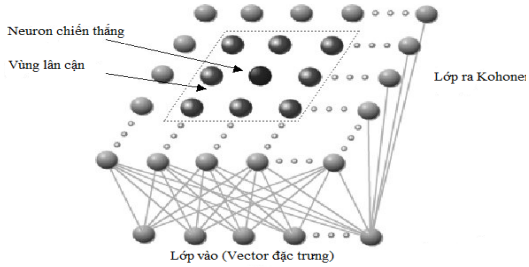
3.1 Mạng Kohonen

Mạng Kohonen do GS. Teuvo Kohonen phát triển vào những năm 1980 [17][27] và đã được ứng dụng vào bài toán gom cụm phẳng. Mạng Kohonen hay còn gọi là phương pháp mạng SOM (Self-Oganizing Maps) được biết đến cho như việc gom cụm dữ liệu mà không cần chỉ định trước số cụm điều này hoàn toàn phù hợp vì không thể xác định trước được số cộng đồng (số cụm) trên mạng xã hội mà số cộng đồng phụ thuộc vào quá trình phân tích chủ đề quan tâm và đề xuất sau quá trình học dữ liệu, ngoài ra mạng Kohonen có khả năng biểu diễn trực quan khối văn bản trên màn hình máy tính thông qua lớp ra Kohonen 2D [12][19][20][22].

Xác định sự phù hợp thông qua nhiều khảo sát các công trình nghiên cứu liên quan, áp dụng phương pháp và giải thuật gom cụm để khám phá cộng đồng mạng theo chủ đề, chúng tôi chọn

phương pháp mạng Kohonen cho hướng nghiên cứu. Mạng Kohonen có thể gom cụm dữ liệu mà không cần chỉ định trước số cụm (tương quan cụm dữ liệu trong nghiên cứu này cộng đồng mạng theo chủ đề, tập ngữ liệu thông điệp vô cùng lớn, đa chiều và cộng đồng mạng rất lớn nên việc xác định trước số cụm - cộng đồng mạng là vô cùng khó khăn)[12][21][22][25]. Ngoài ra, mạng Kohonen có khả năng biểu diễn trực quan khối văn bản, chủ đề thông qua lớp ra Kohonen 2D [12][13][22].

Mục tiêu của mạng Kohonen là ánh xạ những vector đầu vào có N chiều thành một bản đồ với 1 hoặc 2 chiều [12][15][19][20]. Những vector gần nhau trong không gian đầu vào sẽ gần nhau trên bản đồ lớp ra của mạng Kohonen. Một mạng Kohonen bao gồm một lưới các node đầu ra và N node đầu vào.Vector đầu vào được chuyển đến từng node đầu ra (hình 3). Mỗi liên kết giữa đầu vào và đầu ra của mạng Kohonen tương ứng với một trọng số.Tổng đầu vào của mỗi nơron trong lớp Kohonen bằng tổng các trọng của các đầu vào nơron đó.



Hình 3. Cấu trúc của mạng Kohonen[2]

Neuron chiến thắng được xác định bằng cách tìm neuron có khoảng cách ngắn nhất trong tập kết quả. Trong trường hợp này, neuron chiến thắng (winning neuron) là $w_{0,1} = w_{x,y}$. Khi đó ta được: $D_{k1,k2} = D_{0,1} = D_{min}$, với D_{min} = 0.4582. Với k_1 = 0 và k_2 = 1 là chỉ số (dòng, cột) của neuron chiến thắng. Sau khi xác định được neuron chiến thắng, bước tiếp theo xác định vùng lân cận của neuron chiến thắng. Giải thuật sẽ cập nhật lại trọng số của vector trọng của neuron chiến thắng và tất cả các neurons nằm trong vùng lân cận của neuron chiến thắng. Để xác định vùng lân cận của neuron chiến thắng hay gọi là vùng chiến thắng (winning region) ta dùng hàm lân cận (neighborhood function) được áp dụng. Hàm được mô tả như sau:

$$h(r,t) = exp\left(\frac{-r^2}{2\sigma^2(t)}\right) \quad (3)$$

Trong đó, r là khoảng cách từ neuron lân cận đến neuron chiến thắng.

$$r = \sqrt{(k_1 - i)^2 + (k_2 - j)^2} \quad (4)$$

Và $\sigma(t)$: là hàm được sử dụng cho việc xác định không gian lân cận neuron chiến thắng với số lần lặp, giá trị của σ giảm dần [29].

$$\sigma(t) = \sigma_0 e^{-\frac{t}{\tau_1}} \quad (5)$$

Trong đó, (τ_1 là hằng số, $\sigma_0 = \sqrt{m}$, t là số lần lặp). Dưới đây trình bày dạng đơn giản nhất của nhóm hàm mạng lân cận (topological neighborhood function):

$$h(r,t) = \left(1 - \frac{2}{\sigma^2(t)}r^2\right)e^{-\frac{r^2}{\sigma^2(t)}} \quad (6)$$

Áp dụng hàm Mexican để xác định được vùng lân cận neuron chiến thắng cho mỗi vector nhập, trọng số của mỗi neron được cập nhật như sau:

$$w'_{(i,j)_k} = w_{(i,j)_k} + \alpha(t)h(r,t)\left(v_{x_k} - w_{(i,j)_k}\right) \quad (7)$$

$$\forall k \in \mathbb{N}, 0 \le k \le n$$

Trong đó,

- k: chiều của neuron trọng (vector trọng)
- n: số chủ đề được quan tâm
- $w'_{(i,j)_k}$: giá trị mới của neuron trọng thứ k tại dòng i, cột j

[2] http://homepage.ntlworld.com/richard.clark/rs_kohonen.html

- $w_{(i,j)_k}$: giá trị đang xét của neuron trọng thứ k tại dòng i, cột j
- $h(r,t)$: kết quả của hàm mạng lân cận với t số lần lặp, r là khoảng cách giữa neuron đang xét và neuron chiến thắng.
- v_{x_k}: giá trị của vector học v_x thứ k

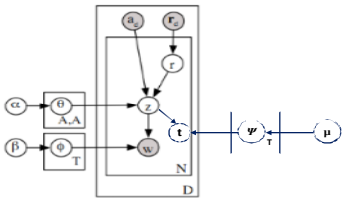
Hàm $\alpha(t)$ là hàm ấn định tốc độ học, giá trị hàm sẽ giảm dần theo số lần lặp t. Nếu một neuron là chiến thắng hay neuron lân cận với neuron chiến thắng, thì trọng của vector đó được cập nhật, ngược lại thì neuron sẽ không được cập nhật. Tại mỗi bước lặp phương pháp Kohonen sẽ quyết định chọn neuron có vector trọng tương tự với vector nhập và điều chỉnh nó và vector trọng lân cận để làm cho chúng gần hơn với vector nhập

Giải thuật 1. Tìm neuron chiến thắng (winning neuron) [19][20][25]

Đầu vào: v, SOM. Trong đó v là vector huấn luyện (vector nhập)
Đầu ra: neuron chiến thắng (winning neuron)
Xử lý:
Bắt đầu
Khởi tạo $min = d(v, SOM[0,0])$;
Khởi tạo $wneuron = SOM[0,0]$;
$Lặp\ i = 0$ đến $sqrt(m)$
$Lặp\ j = 0$ đến $sqrt(m)$
Nếu $min > d(v, SOM[i,j])$
Thì
$min = d(v, SOM[i,j])$;
$wneuron = SOM[i,j]$;
Trả về $neuron$ chiến thắng
Kết thúc

3.2 Mô hình Temporal – Author – Receipent – Topic (TART)

Mô hình TART (hình 4) được chúng tôi đề xuất và giới thiệu trong [25], trong đó có một phần mục tiêu khám phá chủ đề quan tâm của người dùng trên mạng xã hội theo từng giai đoạn thời gian. Cụ thể là tập vector chủ đề quan tâm của người dùng (xem bảng 1 và 2).



Hình 4. Mô hình TART đề xuất [25]

Trong quá trình thực hiện mô hình TART, hệ thống sẽ lưu lại 4 ma trận để phân tích mối quan tâm của người dùng mạng, bao gồm: T (chủ đề) x W (từ), A (tác giả) x T (chủ đề), R (người nhận) x T (chủ đề) and T (chủ đề) x T (thời gian). Dựa trên 4 ma trận, ta có phân bố giữa chủ đề và từ Φ_{zw}, phân bố giữa chủ đề và thời gian Ψ_{zt}, phân bố giữa tác giả và chủ đề Θ_{az}, phân bố giữ người nhận và chủ đề Θ_{rz}. Phân bố của 4 ma trận được xác định bởi biểu thức sau (8), (9), (10) và (11):

$$\theta_{az} = \frac{m_{az} + \alpha}{\sum_z (m_{az} + \alpha)} \quad \phi_{zw} = \frac{n_{zw} + \beta}{\sum_w (n_{zw} + \beta)} \quad (8)\ (9)$$

$$\psi_{zt} = \frac{n_{zt} + \mu}{\sum_t (n_{zt} + \mu)} \quad \theta_{rz} = \frac{m_{rz} + \alpha}{\sum_z (m_{rz} + \alpha)} \quad (10)\ (11)$$

3.3 Mô hình đề xuất tổng quát

Chúng tôi đề xuất mô hình khám phá cộng đồng mạng dựa theo theo mô hình chủ đề có yếu tố thời gian. Trong đó, thông qua kết quả khảo sát, phân tích và đánh giá các mô hình liên quan trọng lĩnh vực khám phá cộng đồng, chúng tôi chọn phương pháp huấn luyện Kohonen; (2) huấn luyện Kohonen kết hợp cải tiến tập dữ liệu đầu vào (là kết quả từ mô hình TART [25]), chính là tập các vector chủ đề quan tâm của người dùng theo từng giai đoạn thời gian. Từ đó, chúng tôi khai thác từng cộng đồng theo các chủ đề quan tâm được thể hiện trên các neurons trên lớp ra Kohonen.

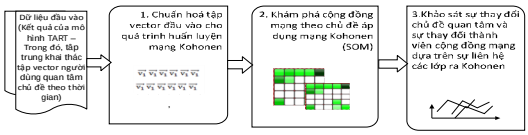
Mô hình thực hiện khám phá cộng đồng thông qua phương pháp gom cụm vector chủ đề

quan tâm của người dùng theo từng giai đoạn thời gian được thực hiện theo mô hình tại hình 5. Mô hình gồm 3 mô-đun chính:

- Chuẩn hoá vector đầu vào: chuẩn hoá dữ liệu đầu vào phù hợp với dữ liệu huấn luyện của mạng Kohonen.

- Khám phá cộng đồng sử dụng mạng Kohonen: áp dụng phương pháp Kohonen để gom cụm người dùng theo chủ đề quan tâm, mỗi cụm là một cộng đồng quan tâm đến các chủ đề và tương ứng với 1 neuron tại lớp ra Kohonen.

- Khảo sát sự thay đổi thành viên và chủ đề quan tâm của cộng đồng dựa trên phân tích sự liên hệ các lớp ra Kohonen.



Hình 5. Mô hình khám phá cộng đồng người dùng theo chủ đề và khảo sát sự thay đổi chủ đề quan tâm và yêu thích của người dùng

Đầu vào: tập vector người dùng quan tâm trao đổi các chủ đề từ kết quả mô hình TART [25]. Thành phần của vector người dùng bao gồm chủ đề mà người dùng quan tâm, xác suất quan tâm và thời gian mà người dùng trao đổi về chủ đề đó.

Đầu ra: tập các cộng đồng người dùng theo các chủ đề cụ thể trong từng khoảng thời gian.

4. KẾT QUẢ THỬ NGHIỆM VÀ THẢO LUẬN

4.1 Dữ liệu thử nghiệm

Thử nghiệm mô hình đề xuất với tập 2055 vector chủ đề quan tâm của 194 người dùng cùng quan tâm trao đổi trên 10 chủ đề (khảo sát ngẫu nhiên trên 10 chủ đề "Cơ sở vật chất và dịch vụ", "Học tập và Thi", "Hợp tác quốc tế", "Kiểm định chất lượng", "Nghiên cứu khoa học", "Sinh hoạt và đời sống", "Thư viện và giáo trình", "Thể dục thể thao", "Tuyển dụng việc làm", "Tuyển sinh", "Tài chính và học phí", "Tình bạn và Tình yêu", "Đoàn hội" và "Đào tạo" trên tổng số 20 chủ đề thuộc hệ thống chủ đề được xây dựng trong [31]). Khảo sát các chủ đề trên trong khoảng thời gian tháng 12-2008 đến tháng 01-2010. Tập vector nhập được xây dựng và chuẩn quá từ kết quả mô hình TART.

Trong từng giai đoạn thời gian, ta sẽ có các vector chủ đề quan tâm của người dùng khác nhau. Chẳng hạn: với người dùng u_1: trong khoảng thời gian từ t_1 đến t_2 vector chủ đề người dùng quan tâm là $v(u_1, t_1, t_{2)}, u \in U$ trong khoảng thời gian t_2 đến t_3 ta có vector $v(u_1, t_2, t_{3)}$.

Một cách tổng quát, mỗi người dùng có một vector chủ đề quan tâm tại thời điểm t là $v_i(t) = < v_{i_1}^t, v_{i_2}^t, v_{i_3}^t, \dots, v_{i_n}^t$. Như vậy, ta có bảng vector chủ đề quan tâm của người dùng như sau:

Bảng 1. Vector quan tâm chủ đề của người dùng

Người dùng	Thời gian t_i	Thời gian t_j	$v(u, t_i, t_{j)}$
u_1	01-11-2008	30-11-2008	$v(u_1, t_1, t_{2)}$
u_2	01-02-2009	28-02-2009	$v(u_2, t_2, t_{3)}$
u_3	01-04-2009	30-04-2009	$v(u_3, t_3, t_{4)}$
u_1	01-02-2009	28-02-2009	$v(u_1, t_2, t_{3)}$

Hay một cách biểu diễn khác về vector chủ đề quan tâm của người dùng:

Bảng 2. Vector quan tâm chủ đề của người dùng

Người dùng	**Chủ đề 1**	**Chủ đề 2**	**Chủ đề 3**	**Thời gian $t_i - t_j$**
	Xác suất quan tâm			
u_1	0.85246	0.0	0.772527	01-11-2008 – 30-11-2008
u_2	0.85000	0.86956	0.676793	01-02-2009 – 28-02-2009
u_3	0.62417	0.34132	0.893421	01-04-2009 – 30-04-2009
u_1	0.52345	0.52341	0.834212	01-02-2009 – 28-02-2009

Bảng 1 và 2 là mẫu các vector chủ đề quan tâm của người dùng trên mạng, đây là tập vector đầu vào cho quá trình huấn luyện mạng Kohonen. Mẫu vector nhập trên bao gồm 3 người dùng quan tâm đến 3 chủ đề trong 3 khoảng thời gian t_1-t_2, t_2-t_3 và t_3-t_4. Mục tiêu quá trình huấn luyện Kohonen là gom cụm các vector chủ đề quan tâm của người dùng.

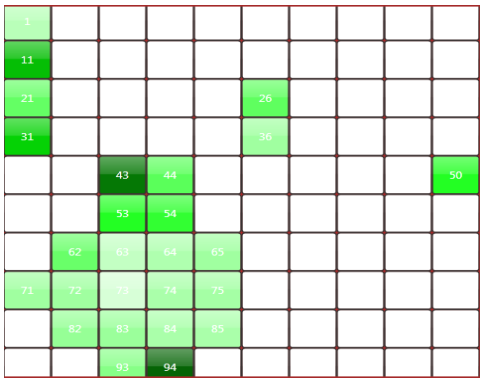
Như vậy, với $V(t_i, t_j)$ ta có lớp ra Kohonen $K(t_i, t_j)$. Đây là mảng 2 chiều (hình 6).Và với tính chất của cụm trên, lớp ra Kohonen ta có danh sách các cụm: {C_1, C_2, C_3, C_4,...,C_k}. Trong đó, mỗi cụm C_i có chứa vector chủ đề của neuron chiến thắng tương ứng.

4.2 Khám phá cộng đồng mạng xã hội

Trong phần này trình bày kết quả thử nghiệm khám phá cộng đồng người dùng trên mạng xã hội theo từng giai đoạn thời gian. Phần này tập trung vào mô-đun (1) và (2) của mô hình tại hình 5.

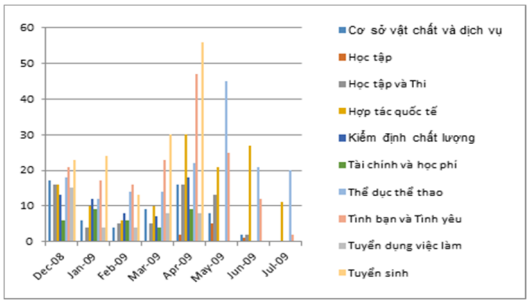
Hình 6 thể hiện kết quả quá trình huấn luyện Kohonen khám phá cộng đồng người dùng mạng theo thời gian với số neuron lớp ra là 100, thử nghiệm trên tập 2055 vector nhập thuộc 194 người dùng quan tâm trao đổi trên 10 chủ đề. Khảo sát trong khoảng thời gian từ 12-2008 đến 01-2010. Số neuron lớp ra được đánh số thứ tự bắt đầu từ 1 cho đến 100 tương ứng 100 neurons (hình 6). Việc xác định số lượng neuron trên lớp ra Kohonen là tuỳ chọn và không làm ảnh hưởng đến kết quả khám phá cộng đồng.

Mỗi neuron lớp ra tương ứng với một cộng đồng những người dùng cùng quan tâm trao đổi chủ đề trong từng giai đoạn thời gian. Với từng neuron, màu sắc đậm và nhạt tương ứng với số lượng người dùng nhiều hay ít tham gia vào cộng đồng. Màu sắc trên mỗi neuron càng đậm đại diện cho số người trong cộng đồng nhiều hơn những neuron có màu nhạt hơn hoặc cộng đồng không có bất kỳ người dùng nào (hiển thị màu trắng) điều này thể hiện cộng đồng không tồn tại.



Hình 6. Kết quả khám phá cộng động được hiển thị bởi tập neurons trên lớp ra Kohonen.

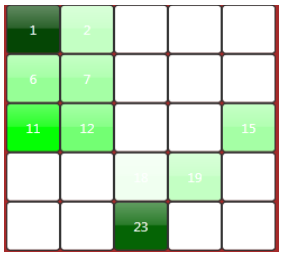
Từ lớp ra Kohonen trên hình 6, chúng tôi tiếp tục khảo sát trình bày kết quả khảo sát sự thay đổi chủ đề và người dùng quan tâmtrong cộng đồng theo từng giai đoạn thời gian (hình 7).



Hình 7. Khảo sát sự thay đổi chủ đề quan tâm và người dùng trong cộng đồng

Hình 7 trình bày kết quả phân tích sự thay đổi chủ đề quan tâm và người dùng trong cộng đồng từ tháng 12/2008 đến tháng 07/2009. Khảo sát trên 10 chủ đề, ta thấy rằng chủ đề mức độ thường xuyên trong các tháng và tăng cao tại các tháng 04, 05/2009 và chiếm đa số người dùng thuộc về các cộng đồng chủ đề"Tuyển sinh", "Thể dục thể thao" và "Tình bạn, tình yêu". Số lượng cộng đồng giảm dần khoảng thời gian tháng 06 và 07/2009. Trong tháng 07/2009 hầu như chỉ có 3 cộng đồng được khám phá, trong đó cộng đồng có số người dùng nhiều nhất là cộng đồng chủ đề "Thể dục thể thao" và ít nhất tại tháng 07/2009 là cộng động chủ đề "Hợp tác quốc tế".

Trên hình 8, tại neuron (cộng đồng) số 23 (vị trí 4, 2) có 80 người dùng quan tâm đến chủ đề Hợp tác quốc tế (hình 8). Đây là cộng đồng có số lượng người dùng đông hơn tất cả các cộng đồng còn lại trong khoảng thời gian khảo sát.

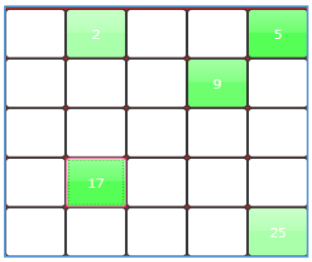


Hình 8. Kết quả khám phá cộng đồng trên lớp ra Kohonen trong khoảng thời gian tháng 04-2009. Mỗi ô hiểu thị trên hình 8 thể hiện một neuron của lớp ra.

Bảng 3. Tập dữ liệu vector chủ đề quan tâm của người dùng trong tháng 12-2008

Người dùng	Chủ đề quan tâm	Thời gian	Số chủ đề quan tâm
U1	Tình bạn và tình yêu	12-2008	3
U1	Đào tạo	12-2008	
U1	Học tập và thi	12-2008	
U3	Thể dục thể thao	12-2008	1
U4	Tình bạn và tình yêu	12-2008	3
U4	Đào tạo	12-2008	
U4	Học tập và thi	12-2008	
U14	Cơ sở vật chất và dịch vụ	12-2008	3
U14	Học tập và Thi	12-2008	
U14	Đào tạo	12-2008	
U20	Đào tạo	12-2008	3
U20	Học tập và Thi	12-2008	
U20	Tình bạn và tình yêu	12-2008	
U36	Tình bạn và tình yêu	12-2008	4
U36	Đào tạo	12-2008	
U36	Học tập và Thi	12-2008	
U36	Thể dục thể thao	12-2008	
U43	Tình bạn và tình yêu	12-2008	1
U49	Đào tạo	12-2008	2
U49	Hợp tác quốc tế	12-2008	
....	.....	.....	.....

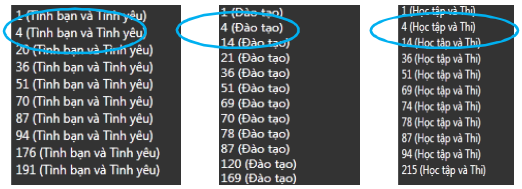
Huấn luyện mạng Kohonen với tập vector chủ đề quan tâm (bảng 3). Ta có lớp ra là tập các neurons (tương ứng mỗi neuron là 1 cụm người dùng theo từng chủ đề đề cụ thể) theo thời gian tháng 12-2008 (hình 9).



Hình 9. Các cộng đồng tham gia trao đổi các chủ đề cụ thể trong tháng 12-2008

Hình 9 chỉ ra rằng, kết quả lớp ra Kohonen gồm có 5 cụm (các neuron có màu). Như vậy, trong tháng 12-2008 có 5 cộng đồng quan tâm đến các chủ đề cụ thể từ tập vector nhập.

Hình 10 thể hiện danh sách các cộng đồng cùng danh sách người dùng tham gia từng chủ đề cụ thể trong tháng 12-2018. Quan sát ta thấy, trong cả 3 cộng đồng người dùng U4 (tương ứng số 4 được khoanh tròn) đều tồn tại. Đều này chứng tỏ, người dùng U4 cùng tham gia vào 3 cộng đồng và quan tâm trao đổi 3 chủ đề cụ thể.



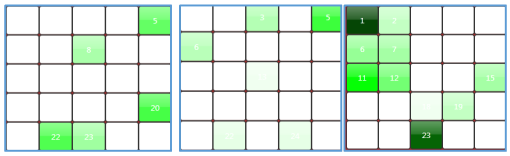
Tại cụm số 25 Tại cụm số 17 Tại cụm số 9

Hình 10. Danh sách các cộng đồng người dùng theo chủ đề quan tâm trong tháng 12-2008 dựa trên lớp ra Kohonen trên hình 9.

4.3 Khảo sát sự thay đổi chủ đề quan tâm và thành viên cộng đồng

Nội dung này tập trung vào phần thử nghiệm mô hình đề xuất của mô-đun (3) tại hình 5. Dựa

trên các lớp ra Kohonen theo từng giai đoạn thời gian, chúng tôi khảo sát được sự liên hệ giữa các cụm (neurons) trên lớp ra Kohonen dựa trên các thành phần của cụm như: người dùng, chủ đề quan tâm, xác suất quan tâm và số cụm hình thành trong từng giai đoạn thời gian.



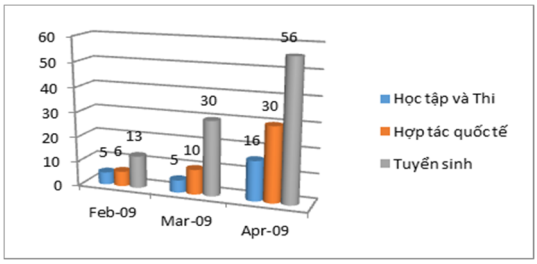
Hình 11a. Cộng đồng tham gia trong tháng 02/2009

Hình 11b. Cộng đồng tham gia trong tháng 03/2009

Hình 11c. Cộng đồng tham gia trong tháng 04/2009

Hình 11. Cộng đồng trên 3 lớp ra Kohonen trong 3 giai đoạn thời gian

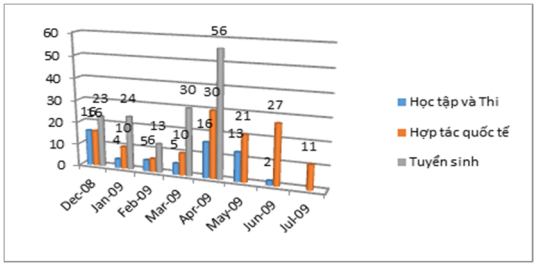
Quan sát trên hình 12, trong tháng 02-2009 có 3 cộng đồng cùng tham gia trao đổi trên mạng. Trong đó, cộng đồng 1 quan tâm đến chủ đề "Hợp tác quốc tế" với số người tham gia lần lượt theo 3 tháng là 6, 10 và 30. Cộng đồng 2 quan tâm đến chủ đề "Tuyển sinh" với số lượng người tham gia lần lượt là 13, 30 và 56. Cộng đồng 3 quan tâm đến chủ đề "Học tập và thi" lần lượt là 5, 5 và 16.



Hình 12. Cộng đồng mạng theo 3 chủ đề trong 3 khoảng thời gian tháng 02, 03 và 04/2009 dựa trên lớp ra Kohonen tại hình 11

Trong từng đơn vị thời gian, mức độ tham gia cộng đồng chủ đề của người dùng mạng cũng có sự thay đổi. Cộng đồng quan tâm đến chủ đề "Tuyển sinh" có số thành viên tham gia lại chiếm ưu thế hơn so với các cộng đồng còn lại.Tuy nhiên, mức độ tham gia vào cộng đồng chủ đề "Học tập và thi" hầu như ít hơn. Điều này phần nào cho thấy trong khoảng thời gian khảo sát trên, việc trao đổi những vấn đề trong học tập, sinh viên rất ít tham gia trao đổi trên mạng xã hội hoặc có những ý kiến về vấn đề học tập.

Quan sát trên hình 13, chúng ta thấy rằng sự co giãn số lượng thành viên trong từng cộng đồng theo từng giai đoạn thời gian. Trong đó, đối với cộng đồng chủ đề "Học tập và thi", thời điểm tháng 12-2008 số thành viên tham gia là 16 nhưng đến tháng 01-2009 số thành viên tham gia cộng đồng này là 4, tháng 06-2009 còn là 2 nhưng đến tháng 07-2009 không tồn tại cộng đồng quan tâm đến chủ đề này. Khảo sát dữ liệu, chúng tôi thấy rằng trong giai đoạn tháng 07-2009 người dùng mạng tham gia trao đổi về chủ đề "Hợp tác quốc tế" là chủ yếu.



Hình 13. Sự thay đổi thành viên cộng đồng mạng xã hội theo chủ đề trong từng giai đoạn thời gian từ tháng 12-2008 đến tháng 07-2009

Tuy nhiên, đến tháng 02-2009 thì số thành viên lại giảm xuống là 4. Đối với cộng đồng quan tâm đến chủ đề "Hợp tác quốc tế", trong tháng 04-2009 có số thành viên tham gia là 24 nhưng đến tháng 05-2009 con số này lại giảm xuống là 4 thành viên. Khảo sát chủ đề "Tuyển sinh" ta thấy đỉnh điểm của cộng đồng chủ đề này là tháng 04-2009 là 56 thành viên tham gia nhưng qua tháng 05, 06 và 07 không còn tồn tại công đồng này. Riêng cộng đồng với chủ đề quan tâm là "Hợp tác quốc tế" tương đối ổn định trong suốt thời gian được khảo sát trên hình 13 từ tháng 12-2008 đến tháng 07-2009.

Như vậy, việc co giãn số lượng thành viên cộng đồng chỉ ra hiện tượng tham gia hoặc rời khỏi cộng đồng của thành viên trong cộng đồng. Nghĩa là tại thời điểm t_i có nhiều hay ít hơn số thành viên trong cộng đồng so với thời điểm t_{i-1} hay t_{i+1}.

4.4 Đánh giá kết quả

Theo Brew C. [26] đã đề nghị phương pháp đánh giá gom cụm như sau: tương ứng với một cụm trong kết quả gom cụm của hệ thống ta tính giá trị của độ đo F-measure với tất cả các cụm được gom bằng tay. Chọn ra giá trị của F-measure cao nhất và loại cụm này ra. Tiếp tục công việc trên, cho các cụm còn lại. Tổng các giá trị F-measure càng cao thì hệ thống gom cụm càng chính xác.

Bảng 4 trình bày kết quả F-measure, **với m = 5 cụm và k =6 cụm.**

Bảng 4. Kết quả tính giá trị F-Measure giữa gom cụm bằng tay (người) và máy

Máy (k) /Người (m)	m_0	m_1	m_2	m_3	m_4
k_0	0.43	0.15	**0.84**	0.52	0.68
k_1	0.67	0.61	0.00	0.16	0.00
k_2	0.00	0.36	0.51	**0.62**	0.16
k_3	0.72	0.00	0.55	0.55	0.34
k_4	**0.81**	**0.73**	0.25	0.00	**0.72**
k_5	0.19	0.00	0.15	0.29	0.36
MAX	**0.81**	**0.73**	**0.84**	**0.62**	**0.72**

Tổng MAX cho gom cụm Kohonen bằng vector: 0.81 + 0.73 + 0.84 + 0.62 + 0.72 = 3.72.

Giá trị tổng max của F-measuare trong bảng 4 là 3.71 tương ứng 74%. Giá trị này theo chúng tôi đánh giá là cao, điều này chứng tỏ phương pháp đề xuất gom cụm người dùng bằng phương pháp mạng Kohonen dựa trên tập vector chủ đề quan tâm theo thời gian có độ chính xác cao.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Đóng góp trong nghiên cứu này được tổng hợp thành 2 nội dung chính:

1. Đề xuất mô hình khám phá cộng đồng mạng xã hội dựa theo mô hình chủ đề có yếu tố thời gian. Mô hình đề xuất không cần phải xác định trước số cộng đồng (số cụm) điều này hoàn toàn phù hợp với tính chất của mạng xã hội không thể biết được số lượng cộng đồng đang tồn tại và cộng đồng thì thường xuyên thay đổi.

Trong đó, chúng tôi tập trung khai thác và kết hợp phương pháp mạng Kohonen kết hợp mô hình TART [25]. Phương pháp thực hiện gồm 2 phần chính: (1) chuẩn hoá và chọn kết quả là tập các vector chủ đề quan tâm của từng người dùng trên mạng xã hội, đây chính là tập vector đầu vào của quá trình huấn luyện mạng Kohonen, (2) đề xuất mô hình áp dụng phương pháp huấn luyện mạng Kohonen để khám phá cộng đồng những người dùng cùng quan tâm đến từng chủ đề cụ thể được gọi là cộng đồng mạng theo chủ đề. Trong đó, mô hình có thể khám phá được chủ đề theo từng giai đoạn thời gian được cộng đồng mạng quan tâm, mức độ quan tâm; tính được phân bố chủ đề theo từng cộng đồng mạng. Thách thức đặc ra trong nghiên cứu này là khám phá cộng đồng theo chủ đề dựa trên nội dung trao đổi trên mạng xã hội bởi gì cộng đồng thường xuyên thay đổi chủ đề quan tâm cũng như thay đổi thành viên tham gia cộng đồng mạng xã hội.

2. Khảo sát sự thay đổi chủ đề quan tâm và người dùng trong cộng đồng mạng xã hội theo từng giai đoạn thời gian dựa trên sự liên hệ các lớp ra Kohonen. Điều này giúp cho việc theo dõi sự thay đổi sự quan tâm của người dùng trên mạng xã hội chịu ảnh hưởng của sự thay đổi chủ đề quan tâm của cộng đồng mà người dùng đó tham gia.

5.2 Hướng phát triển

Kết quả bài báo nghiên cứu này này sẽ là nền tảng cho những nghiên cứu tiếp theo sau này như tìm kiếm người quan trọng trong cộng đồng mạng, phân tích ảnh hưởng lan truyền chủ đề và tìm kiếm nguồn gốc của thông tin trên mạng xã hội.

Lời cám ơn: *Nghiên cứu này được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM) trong đề tài mã số B2013-26-02.*

A New Model for Discovering Communities of Users on Social Network

- **Thanh Ho** [1]
- **Phuc Do** [2]

[1] Faculty of Information System, University of Economics and Law, VNU-HCM

[2] University of Information Technology, VNU-HCM

ABSTRACT

The trend of technological development and increasing varieties of social media lead to the changes in people's behaviors in society and forming online communities. Changes of human's behaviors make many models of business, marketing, services and even the field of education, security, politicsl change from approaches to user management. Community of users on social networks influence behaviors, habits of each user involved in the community. Therefore, exploring community on social networks from many different data sources via analyzing exchanged contents will help know the user community's behaviors which are reflected in the content and topics that users are interested in discussing in messages. In this paper, we propose a new model of discovering communities of users on social networks based on the topic model combined with Kohonen network. In the proposed model, we focus on discovering communities of users on social networks and analyzing the interested topics change of online community in each period of time. The proposed model is experimented with a set of vectors in interested topics of online users in higher education field.

Keywords: *topic, topic model, discovering communities, analyzing changes, Kohonen Network, TART.*

TÀI LIỆU THAM KHẢO

[1]. X. Wang, N. Mohanty, and A. McCallum (2006). Group and topic discovery from relations and their attributes. Advances in Neural Information Processing Systems 18, pp. 1449-1456.

[2]. N. Pathak, C. DeLong, A. Banerjee, and K. Erickson (2008), Social topic models for community extraction. In The 2nd SNA-KDD Workshop, volume 8.

[3]. D. Zhou, E. Manavoglu, J. Li, C.L. Giles, and H. Zha (2006), Probabilistic models for discovering e-communities. In WWW '06: Proceedings of the 15th international conference on World Wide Web, page 182. ACM, pp. 173-182.

[4]. István Bíró, Jácint Szabó (2008), Latent Dirichlet Allocation for Automatic Document Categorization, Research Institute of the Hungarian Academy of Sciences Budapest, pp. 430-441.

[5]. Andrew McCallum, Andr´es Corrada, Xuerui Wang (2004), The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email, Department of Computer Science, University of MA.

[6]. Michal Rosen-Zvi, Thomas Griffths et. al (2004), Probabilistic AuthorTopic Models for Information Discovery, 10th ACM SigKDD, Seattle, pp. 306-315.

[7]. Alexandru Berlea1, Markus Döhring, Nicolai Reuschling (2009), Content and communication based sub-community detection using probabilistic topic models, IADIS International Conference Intelligent Systems and Agents.

[8]. Wenjun Zhou, Hongxia Jin, Yan Liu (2012), Community Discovery and Profiling with Social Messages, KDD'12, August 12–16, 2012, Beijing, China, pp. 388-396.

[9]. Chunshan Li, William K. Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, Xin Li (2014), The Author-Topic-Community model for author interest profiling and community discovery, Springer-Verlag London 2014, pp. 74-85.

[10]. The Anh Dang, Emmanuel Viennet (2012), Community Detection based on Structural and Attribute Similarities, ICDS 2012 : The Sixth International Conference on Digital Society, pp. 7-14.

[11]. Yang Zhou, Hong Cheng, Jeffrey Xu Yu (2009), Graph Clustering Based on Structural/Attribute Similarities, VLDB '09, August 24-28, 2009, Lyon, France, pp. 718-729.

[12]. Do Phuc, Mai Xuan Hung (2008), Using SOM based Graph Clustering for Extracting Main Ideas from Documents, RVIF 2008, pp. 209-214.

[13]. Kohonen T. and Honkela T. (2007), Kohonen network, http://www.scholarpedia.org/article/Kohonen_network.

[14]. Zhijun Yin et. al (2012), Latent community Topic Analysis: Integration of Community Discovery with Topic Modeling, ACM Transactions on Intelligent Systems and Technology, pp. 1-21.

[15]. Kaski, S., Honkela, T., Lagus, K., and Kohonen. T.WEBSOM--self-organizing maps of document collections. Neurocomputing, volume 21, (1998), pp. 101-117.

[16]. Thanh Ho, Phuc Do (2015), Analyzing Users' Interests with the Temporal Factor Based on Topic Modeling, 23-25 March 2015, Indonesia, Springer, pp. 106-115.

[17]. Teuvo Kohonen (1982), Self-Organized Formation of Topologically Correct Feature Maps, Biol. Cybern. 43, Springer-Verlag,npp. 59-69.

[18]. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps.*Biological Cybernetics*, 43:59-69

[19]. Kohonen T. (1984). *Self-Organization and Associative Memory*, Springer, Berlin.

[20]. Kohonen, T. (2001). *Self-Organizing Maps*. Extended edition. Springer.

[21]. Kohonen, T., Kaski, S. and Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, 9: 1321-1344.

[22]. Kohonen, T. and Somervuo, P. (2002). How to make large self-organizing maps for nonvectorial data. *Neural Networks* 15(8-9), pp. 945-952.

[23]. Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, Rong Jin (2011), Detecting communities and their evolutions in dynamic social networks—a Bayesian approach, Mach Learn 82, Springer, pp. 157–189.

[24]. Ding Zhou, Isaac Councill, Hongyuan Zha, C. Lee Giles (2007), Discovering Temporal Communities from Social Network Documents, IEEE ICDM, pp. 745-750.

[25]. Tran Quang Hoa, Vo Ho Tien Hung, Nguyen Le Hoang, Ho Trung Thanh, Do Phuc (2014), Finding the Cluster of Actors in Social Network based on the Topic of Messages, ACIIDS 04/2014, ThaiLan. Springer, pp. 183-190.

[26]. Brew C, Schulte im Walde (2002). Spectral Clustering for German Verbs, In Proc of the Conf in Natural Language Proocessing, Philadenphia, PA, pp. 117-124.

[27]. Yan Liu, Alexandru N.M et al (2009), Topic-Link LDA: Joint Models of Topic and Author Community, Proceedings of the 26 th International Conference on Machine Learning, ACM, pp. 665-672.

[28]. Mr inmaya Sachan, et al (2012), Using Content and Interactions for Discovering Communities in Social Networks, International World Wide Web Conference Com-mittee (IW3C2), Lyon, France, pp. 331-340.

[29]. B. Magomedov, "Self-Organizing Feature Maps (Kohonen maps)," 7 November 2006. [Online]. Available: http://www.codeproject.com/Articles/16273/Self-Organizing-Feature-Maps-Kohonen-maps.

[30]. Nguyen Le Hoang, Do Phuc, et al (2013), Predicting Preferred Topics of Authors based on Co-Authorship Network, The 10th IEEE RIVF International Conference on Computing and Communication Technologies, IEEE, pp. 70-75.

[31]. Hồ Trung Thành, Đỗ Phúc (2014), *Ontology tiếng Việt trong lĩnh vực giáo dục đại học,* Tạpchí Khoa học Công nghệ, Viện Hàn lâm Khoa học Công nghệ Việt Nam, Tập 52, số 1B, pp. 89-100.

[32]. Tom Fawcett (2005), Introduction to ROC Analysis, Elsevier B.V., Available online www.sciencedirct.com