Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý 3(3):311-326

Open Access Full Text Article

Bài nghiên cứu

Khám phá và trực quan hoá cộng đồng cá nhân trên mạng xã hội dựa theo mô hình chủ đề kết hợp mạng Kohonen

Hồ Trung Thành*, Nguyễn Quang Hưng, Trần Duy Thanh

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Cá nhân (người dùng) là thành viên của cộng đồng trên mạng xã hội. Chủ đề quan tâm của cá nhân trên mạng xã hội thường thay đổi dẫn đến chủ đề quan tâm của cộng đồng thay đổi theo. Mức độ, thời gian và chủ đề được quan tâm của cá nhân trong cộng đồng là những đặc trưng của cộng đồng. Sự thay đổi các đặc trưng của cộng đồng thường phụ thuộc vào hai nguyên nhân chính : (i) thông qua sở thích của từng cá nhân trên mạng cùng kết bạn với nhau hoặc cùng quan tâm đến những chủ đề dựa trên nội dung thông điệp mà cá nhân quan tâm trao đổi ; (ii) hình thành hay thay đổi từ nhóm các bạn bè cùng kết bạn trên mạng hoặc thông qua sự giới thiệu bạn bè cùng kết bạn. Như vậy, mối liên hệ của cá nhân trong cộng đồng xem như một mạng liên kết những thành viên thông qua những đặc trưng trên MXH. Trong bài báo này, tác giả nghiên cứu và đề xuất phương pháp khám phá cộng đồng sử dụng mô hình chủ đề có yếu tố thời gian TART kết hợp phương pháp mạng nơ-ron Kohonen với mục tiêu khám phá cộng đồng những cá nhân có cùng chủ đề quan tâm theo từng giai đoạn thời gian. Qua thử nghiệm mô hình và phương pháp được đề xuất trên hai tập dữ liệu thông điệp tiếng Việt (thu thập từ mạng xã hội trong các trường đại học và trang báo điện tử) bằng hệ thống phần mềm được xây dựng để phân tích mạng mạng xã hội đã đạt được mục tiêu của nghiên cứu.

Từ khoá: khám phá cộng đồng, phân tích mạng xã hội, mô hình TART, mạng nơ-ron Kohonen, mô hình chủ đề

GIỚI THIỆU

Mạng xã hội trực tuyến (MXH) đã đạt được những thành tựu lớn trong nhiều lĩnh vực như kinh tế, chính trị, xã hội, giáo dục. Mục tiêu phân tích MXH là phân tích sự tương tác giữa con người, tổ chức với nhau và khám phá những thông tin, tri thức tiềm ẩn thông qua sự tương tác đó [1–4]. Xu hướng gần đây, các nghiên cứu tập trung vào phân tích MXH và khám phá cộng đồng. Chính MXH đã tạo nên sự không lệ thuộc vào không gian và thời gian khi giao tiếp của cá nhân và cộng đồng. MXH mang lại lượng lớn dữ liệu là thông điệp trao đổi của cá nhân thông qua các liên kết xã hội. **Hình 1** biểu diễn mối liên kết giữa các cá nhân trong MXH.

Có thể biểu diễn dữ liệu này bằng cấu trúc đồ thị của MXH và nội dung dữ liệu là thông tin trao đổi giữa các thành viên trên MXH trong đó bao gồm dữ liệu thông điệp, dữ liệu đa phương tiện... Đây chính là nguồn dữ liệu để phân tích MXH tìm ra những thông tin, tri thức tiềm ẩn được chứa đựng trong dữ liệu trên MXH [2,3,5].

Khám phá cộng đồng là một phương pháp trong phân tích MXH nhằm tìm ra các nhóm những cá nhân có mối liên kết xã hội với nhau trên MXH và cùng chủ đề quan tâm [6–10], đồng thời giúp hiểu được sự quan tâm của từng cá nhân trong cộng đồng MXH theo từng chủ đề cụ thể. Những thay đổi xảy ra trong cộng đồng thường liên quan đến các đặc trưng của cộng đồng như : chủ đề quan tâm, số cá nhân tham gia cộng đồng, mức độ quan tâm chủ đề của cộng đồng tại từng thời điểm khác nhau, và sự thay đổi chủ đề quan tâm trong cộng đồng dẫn đến thay đổi hành vi, sự quan tâm và trao đổi chủ đề của các cá nhân trong cộng đồng.

Nghiên cứu đặt ra là làm thế nào để có thể khám phá cộng đồng cùng quan tâm đến một hay một nhóm chủ đề thông qua những nội dung thông điệp được trao đổi bởi các cá nhân trên MXH? Với một hay nhóm chủ đề cụ thể có những cộng đồng nào trên MXH quan tâm trao đổi ? Sự biến thiên chủ đề quan tâm và cá nhân tham gia cộng đồng ? Tìm giải pháp cho các câu hỏi này rõ ràng là việc không đơn giản nhưng kết quả nghiên cứu sẽ giúp cho việc phân tích và khám phá chủ đề được cá nhân quan tâm hay tìm ra những cá nhân có ảnh hưởng trong cộng đồng để phục vụ cho những chiến lược phát triển như quản lý cộng đồng cá nhân của công ty, tổ chức hay của một quốc gia; hiểu cá nhân để thực hiện chiến lược tiếp thị hiệu quả, phát

Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Liên hệ

Hồ Trung Thành, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Email: thanhht@uel.edu.vn

Lịch sử

- Ngày nhận: 19/2/2019
- Ngày chấp nhận: 25/4/2019
- Ngày đăng: 30/9/ 2019

DOI :10.32508/stdjelm.v3i3.572

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Thành H T, Quang Hưng N, Duy Thanh T. **Khám phá và trực quan hoá cộng đồng cá nhân trên mạng xã hội dựa theo mô hình chủ đề kết hợp mạng Kohonen**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 3(3):311-326.

Hình 1: **Mối liên kết xã hội giữa các cá nhân (actors) trên MXH Facebook.** *(Nguồn: http://www.sangnghiep.com)*

triển loại hình đào tạo trực tuyến trong trường đại học và ứng dụng trong nhiều lĩnh vực khác.

CÁC NGHIÊN CỨU LIÊN QUAN

Bài nghiên cứu tập trung khảo sát các nghiên cứu về xây dựng mô hình khám phá nhóm hay cộng đồng cá nhân trên MXH cùng quan tâm đến chủ đề[9,11–14]. Bên cạnh đó, bài nghiên cứu cũng đã khảo sát các nghiên cứu liên quan đến khám phá cộng đồng MXH[1,12,15–19] dựa theo mô hình chủ đề. Các nghiên cứu trên đã đạt kết quả trong khám phá cộng đồng mạng dựa trên việc phân tích nội dung thông điệp là các bài báo khoa học, nội dung email bằng tiếng Anh. Trong đó, một số mô hình tiêu biểu như mô hình GT (Group – Topic)[13] được xây dựng dựa theo phương pháp mạng Bayes, mục tiêu của mô hình GT là khám phá những nhóm cá nhân ẩn trên MXH dựa trên phân tích nội dung được trao đổi bởi cá nhân. Tuy nhiên, nghiên cứu này chưa chỉ rõ từng thành phần trong cộng đồng như cá nhân gửi, cá nhân nhận thông điệp. Mô hình CUT (Community-User-Topic)[8] đã dựa theo phương pháp mạng Bayes, kỹ thuật Gibbs sampling và phương pháp khám phá cộng đồng để tìm ra tập cá nhân cùng quan tâm đến các chủ đề cụ thể và hình thành nên các cộng đồng. Tuy nhiên, trong tài liệu của Zhou và cộng sự[8] giống như một số mô hình khác đã giới thiệu, Zhou và cộng sự[8] chưa quan tâm đến yếu tố thời gian mà cá nhân hay cộng đồng quan tâm trao đổi chủ đề cũng chưa quan tâm đến cá nhân là người nhận hay người gửi trong cộng đồng. Việc này là quan trọng để phân tích được xu thế quan tâm chủ đề với vai trò của cá nhân. Mô hình ATC (Author-Topic-Community)[7] được nhóm tác giả đề xuất và công bố vào năm 2015. Mô hình ATC tập trung quan tâm khai thác các thành phần chính là tác giả A, cộng đồng C và chủ đề T. Trong nghiên cứu[7], nhóm tác giả chưa tập trung khai thác yếu tố thời gian và phân tích sự biến thiên chủ đề quan tâm của cộng đồng cũng như cá nhân trên MXH.

ĐỘNG LỰC NGHIÊN CỨU

Đối với các nghiên cứu được giới thiệu trên, chúng ta nhận thấy rằng:

- Ưu điểm:

- Các mô hình đã được xây dựng dựa theo mô hình chủ đề.
- Sử dụng mô hình ART[20] để tạo vector chủ đề quan tâm và sử dụng làm vector đầu vào cho quá trình gom cụm của mô hình.
- Các mô hình dùng giải thuật gom cụm (K-Means hoặc K-Medoids, và một số giải thuật khác) để khám phá cộng đồng MXH theo vector chủ đề quan tâm.

- Hạn chế:

- Chưa gom cụm cộng đồng theo thời gian vì vector đầu vào của mô hình ART[20] không có yếu tố thời gian.
- Chưa biểu diễn trực quan kết quả gom cụm cộng đồng theo thời gian và phân tích sự biến thiên đặc trưng của cộng đồng.
- Số cộng đồng MXH là rất lớn, các nghiên cứu dùng giải thuật K-Means hoặc K-Medoids nên khó tính toán trước hệ số K để gom cụm cộng đồng. Nghĩa là khó xác định số cộng đồng.

Bên cạnh đó, đối với vấn đề phân tích sự phân bố chủ đề trong cộng đồng theo thời gian, phân bố chủ đề được quan tâm trong cộng đồng, với một chủ đề thì sự quan tâm của nhiều cá nhân 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 dựa trên tập dữ liệu thông điệp tiếng Anh. Bài báo nghiên cứu và thử nghiệm trên tập dữ liệu thông điệp tiếng Việt được thu thập từ MXH.

Để khắc phục những hạn chế của các nghiên cứu trước, bài nghiên cứu xây dựng phương pháp khám phá cộng đồng dựa trên mô hình chủ đề có yếu tố thời gian kết hợp mạng nơ-ron Kohonen để khám phá cộng đồng theo thời gian và trực quan hoá kết quả khám phá cộng đồng dựa trên lớp ra Kohonen. Áp dụng phương pháp huấn luyện Kohonen để gom cụm những cá nhân cùng quan tâm đến chủ đề cụ thể những mức độ quan tâm là khác nhau, vì thế kết quả gom nhóm từ phương pháp đề xuất của bài nghiên cứu giải quyết được tiêu chí phải xác định trước số cụm trong phương pháp gom cụm.

LÝ THUYẾT KHÁM PHÁ CỘNG ĐỒNG CÁ NHÂN TRÊN MẠNG XÃ HỘI

Theo **Hình 2** và **Hình 3** thể hiện một mô hình MXH gồm các cộng đồng cá nhân[9].

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 ta có $c \in C$[9].

Định nghĩa 1:Cộng đồng[9]

Cộng đồng là một tập thể cùng sống và làm việc trong cùng một môi trường.

Định nghĩa 2:Cộng đồng MXH[5,9]

Cộng đồng MXH 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 **Hình 2**.

Định nghĩa 3: Cộng đồng MXH theo chủ đề (đề xuất của bài nghiên cứu)

Dựa theo mô hình chủ đề, cộng đồng là tập hợp các cá nhân cùng quan tâm đến các chủ đề. Mỗi cá nhân trong cộng đồng được đặc trưng bằng một vector chủ đề quan tâm và mức độ cùng quan tâm đến chủ đề trong cộng đồng nhiều hơn so với những cộng đồng khác. Cho c là một cộng đồng theo chủ đề, $c \in C$, trong đó C là tập hợp các cộng đồng. Cộng đồng là một phân hoạch với các đặc tính như cụm, ký hiệu $C = \{C_l, C_2, C_3, C_4, \ldots, C_K\}$ với K là số cộng đồng, mỗi cộng đồng C_i có tập vector chủ đề:

1. Rời nhau: $C_i \cap C_j = \emptyset$ nếu hai cộng đồng không cùng quan tâm đến một hay nhiều chủ đề cụ thể nào đó (**Hình 3**).
2. Và hợp các cộng đồng $U_{i=1}^{K} C_i = C$

Định nghĩa 3 được bài nghiên cứu áp dụng để thử nghiệm phương pháp khám phá cộng đồng.

PHƯƠNG PHÁP GOM CỤM, VẤN ĐỀ TRỰC QUAN HÓA VÀ MÔ HÌNH CHỦ ĐỀ

Phương pháp gom cụm và vấn đề trực quan hóa

Phương pháp gom cụm (khám phá cộng đồng) là quá trình nhận biết các cụm dữ liệu mà mỗi cụm là một tập hợp dữ liệu tương đồng nhau. Sự tương đồng nhau của dữ liệu được mô tả và xác định bởi hàm khoảng cách tuỳ thuộc vào từng phương pháp (thường là khoảng cách Euclide). Mục đích gom cụm dữ liệu cũng nhằm nhận diện mật độ dữ liệu trong tập dữ liệu lớn, nhiều chiều từ đó hiểu được cấu trúc của dữ liệu đầu vào và nhận biết những cụm dữ liệu có những đặc trưng giống nhau. Có nhiều kỹ thuật gom cụm dữ liệu như SVM, K-means, K-Medoids và mạng nơ-ron Kohonen (hay còn gọi là Self-Organizing Map (SOM)[21]. Mạng nơ-ron Kohonen do GS. Teuvo Kohonen phát triển vào những năm 1980[21] và đã được ứng dụng vào bài toán gom cụm phẳng. Mạng nơ-ron Kohonen gom cụm dữ liệu mà không cần chỉ định trước số cụm. Điều này tương quan với cụm dữ liệu trong nghiên cứu này là cộng đồng mạng theo chủ đề, tập dữ 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. Một mục tiêu quan trọng của mạng nơ-ron Kohonen đối với nghiên cứu này là khả năng biểu diễn trực quan kết quả khám phá cộng đồng trên lớp ra Kohonen 2D[21].

Mục tiêu cụ thể của mạng nơ-ron Kohonen là ánh xạ những vector đầu vào có N chiều thành một bản đồ với 1 hoặc 2 chiều[21–23]. Những vector gần nhau trong không gian đầu vào sẽ gần nhau trên bản đồ lớp ra của mạng nơ-ron Kohonen. Điều này đã giúp bài nghiên cứu giải quyết được vấn đề đưa vector chủ đề quan tâm của cá nhân (kết quả mô hình TART[24]) nhiều chiều về vector 2 chiều để trực quan hóa trên lớp ra mạng nơ-ron Kohonen.

Một mạng nơ-ron Kohonen bao gồm một lưới các node đầu ra và N node đầu vào. Vector đầu vào được chuyển đến từng node đầu ra. Mỗi liên kết giữa đầu vào và đầu ra của mạng nơ-ron Kohonen tương ứng với một trọng số. Theo tính chất của thuật giải huấn luyện trên mạng nơ-ron Kohonen, các cụm có vị trí gần nhau trên mạng nơ-ron Kohonen sẽ chứa các đối tượng có mức độ tương tự cao (tập văn bản có nội dung tương tự nhau).

Hình 2: C**ác cộng đồng có liên hệ trong MXH.***(Nguồn : http://www.smartinsights.com)*

Hình 3: **Các cộng đồng rời rạc trong MXH.***(Nguồn: http://www.website-building-and-hosting.com)*

Mô hình chủ đề có yếu tố thời gian TART

Mô hình TART (Temporal-Author-Recipient-Topic) phân tích MXH có yếu tố thời gian dựa theo mô hình chủ đề (**Hình 4**).

Nhiệm vụ của mô hình TART[24]:

- Khám phá chủ đề quan tâm của cá nhân trên MXH có yếu tố thời gian. Nghĩa là tìm tập actor vector có yếu tố thời gian.
- Phân tích vai trò của cá nhân tham gia mạng xã hội dựa theo mô hình chủ đề có yếu tố thời gian.
- Dùng yếu tố thời gian để chia nhỏ các yếu tố trong mô hình ART như tập cá nhân gửi, tập cá nhân nhận, tập chủ đề và tìm ra sự thay đổi chủ đề quan tâm của cá nhân trong tập thông điệp theo từng khoảng thời gian.
- Khảo sát sự biến thiên chủ đề quan tâm của từng cá nhân.

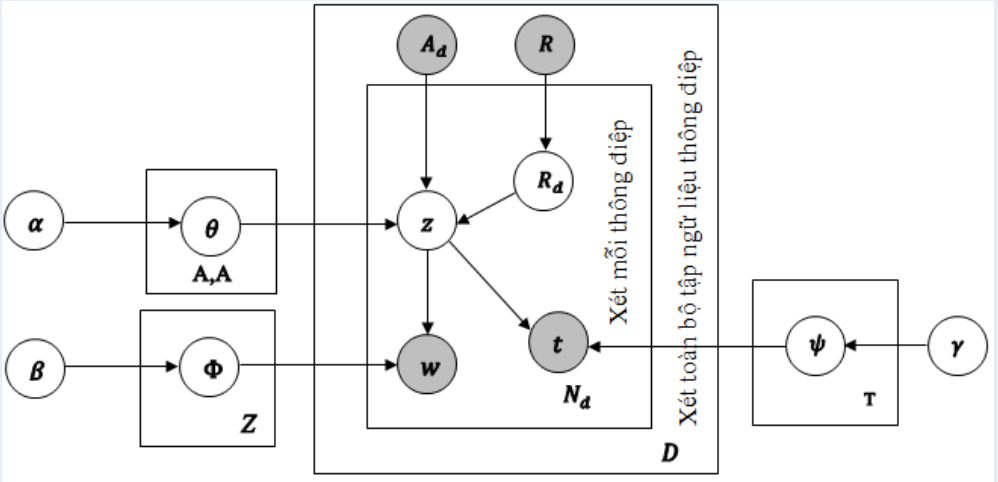
ĐỀ XUẤT PHƯƠNG PHÁP KHÁM PHÁ CỘNG ĐỒNG

Phương pháp khám phá cộng đồng

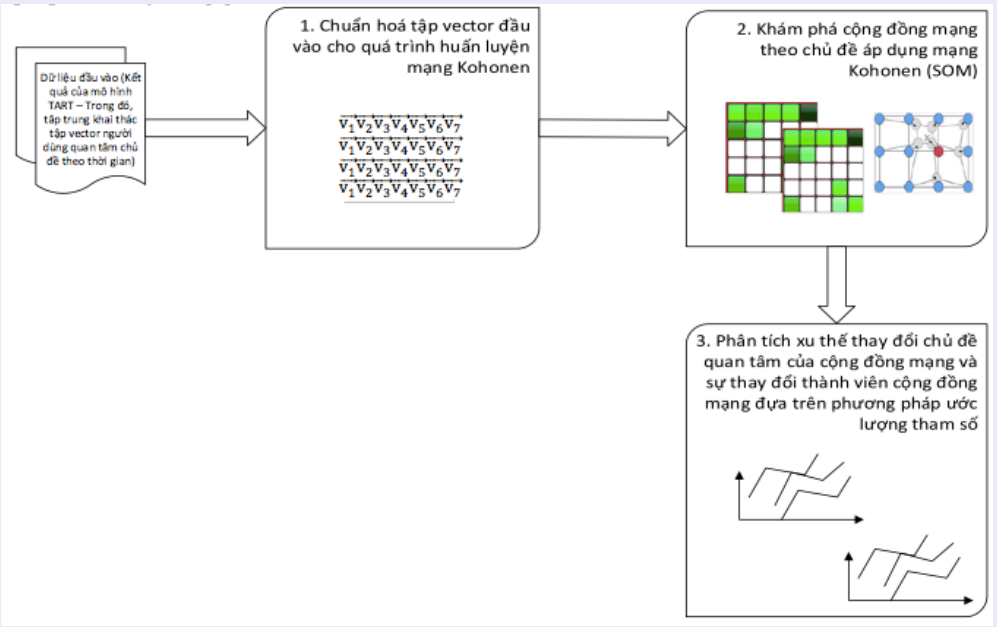
Phương pháp khám phá cộng đồng cá nhân trên MXH dựa theo mô hình chủ đề để khám phá cộng đồng được đề xuất với 2 nhiệm vụ chính: (i) xây dựng phương pháp khám phá cộng đồng dựa theo mô hình chủ đề có yếu tố thời gian. Trong đó, thông qua kết quả khảo sát, phân tích và đánh giá các mô hình liên quan khám phá cộng đồng, bài nghiên cứu chọn phương pháp huấn luyện Kohonen ; (ii) huấn luyện mạng nơ-ron Kohonen kết hợp chuẩn hóa tập dữ liệu đầu vào (là kết quả được thực hiện từ mô hình TART) là tập các vector chủ đề quan tâm của cá nhân theo từng giai đoạn thời gian. Từ đó, bài nghiên cứu thực hiện phương pháp khám phá cộng đồng cá nhân và kết quả được thể hiện trên các nơ-ron của lớp ra Kohonen.

Phương pháp khám phá cộng đồng thông qua phương pháp gom cụm dựa trên vector đặc trưng của cá nhân theo từng giai đoạn thời gian. Phương pháp này được thực hiện như **Hình 5**. Phương pháp được xây dựng gồm ba mô-đun chính:

i. Chuẩn hoá vector nhập: là việc chuẩn hoá dữ liệu đầu vào phù hợp với dữ liệu huấn luyện của mạng nơ-ron Kohonen. Chuẩn hoá vector nhập cho quá trình huấn luyện mạng nơ-ron Kohonen là cần thiết[23]. Cụ thể Mô-đun này thực hiện việc chuẩn hoá tập vector chủ đề quan tâm của cá nhân theo từng giai đoạn thời gian từ kết quả mô hình TART thành tập vector nhập cho huấn luyện mạng nơ-ron Kohonen. Bởi vì các thành phần vector chủ đề quan tâm của mô hình



Hình 4: **Mô hình TART khám phá chủ đề quan tâm của cá nhân theo thời gian**[24]



Hình 5: **Phương pháp khám phá cộng đồng dựa theo chủ đề và phân tích sự biến thiên chủ đề quan tâm của cá nhân.**

TART có thể cho giá trị > 1. Điều này không thoả điều kiện không gian vector của vector trọng nằm trong khoảng [0,1].

ii. Khám phá cộng đồng sử dụng mạng nơ-ron Kohonen: áp dụng mạng nơ-ron Kohonen để gom cụm cá nhân theo chủ đề quan tâm, mỗi cụm là một cộng đồng theo chủ đề và tương ứng với 1 nơ-ron tại lớp ra Kohonen.

iii. Phân tích sự biến thiên đặc trưng của cộng đồng: dựa trên lớp ra Kohonen phân tích sự biến thiên cá nhân tham gia cộng đồng và chủ đề mà cộng đồng quan tâm theo từng giai đoạn thời gian.

Phát biểu bài toán khám phá chủ đề quan tâm của cộng đồng trên MXH

Áp dụng mạng nơ-ron Kohonen để gom cụm cá nhân theo chủ đề quan tâm. Dựa trên tập vector chủ đề

quan tâm của cá nhân theo từng giai thời gian, quá trình huấn luyện để gom cụm dựa trên vector đặc trưng từ mô hình TART[24]. Mỗi cụm là một cộng đồng cá nhân cùng quan tâm đến nhiều chủ đề theo từng giai đoạn thời gian và được hiển thị trên mỗi nơ-ron tại lớp ra Kohonen.

Cho MXH $G = < V,E >$, trong đó V là tập các cá nhân, E là tập các thông điệp trao đổi giữa các cá nhân và cho tập vector chủ đề quan tâm của cá nhân, tìm cộng đồng C gồm các cá nhân có cùng chủ đề và mức độ quan tâm chủ đề theo từng giai đoạn thời gian.

Cho: tập vector nhập (vector chủ đề quan tâm của cá nhân) $\{v_i\}$ là kết quả từ mô hình TART. Vector v_i có m chiều, $v_i < v_{i1}, v_{i2}, \ldots, v_{im} > m$ là số chủ đề quan tâm. Thành phần của vector nhập bao gồm tập chủ đề mà cá nhân quan tâm, mức độ quan tâm và thời gian cá nhân quan tâm chủ đề.

Tìm: danh sách các cộng đồng cá nhân $C = \{C_1, C_2, C_3, C_4, \ldots, C_K\}$ quan tâm đến tập chủ đề theo từng giai đoạn thời gian. Đặc trưng của từng cộng đồng C_i bao gồm: chủ đề quan tâm, mức độ quan tâm chủ đề và số cá nhân tham gia cộng đồng. Với K là số cộng đồng. Trong đó, các cộng đồng có tính chất:

• Rời rạc nhau: $C_i \cap C_j = \emptyset$ nếu các cộng đồng không cùng quan tâm đến một hay nhiều chủ đề cụ thể nào đó.

• Và hợp các cộng đồng $U_{i=1}^{K} C_i = C$.

Phương pháp: áp dụng mạng nơ-ron Kohonen[21,22], các bước xử lý chính sau:

i. Chuẩn hóa vector nhập v_i

ii. Đưa vector nhập v_i vào quá trình huấn luyện mạng nơ-ron Kohonen

iii. For each $i \in [\ 1,...,n\ ]$ //n là số cột và dòng lớp ra Kohonen

For each $j \in [\ 1,...,n\ ]$

Tìm nơ-ron có vector trọng w_{ij} gần với vector nhập v nhất

Gọi (i_0 , j_0) là toạ độ của nơ-ron chiến thắng. Như vậy, khoảng cách $d\left(v, w_{i0,j0}\right) = min\left(d\left(v, w_{ij}\right)\right)$, với $i, j \in [1, \ldots, n]$ và $w_{i0,j0}$ là trọng của nơ-ron chiến thắng.

iv. Xác định vùng lân cận và cập nhật nơ-ron chiến thắng (xem **Hình 6**).

Mạng SOM áp dụng cạnh tranh mềm để gom cụm dữ liệu. Vì vậy, không những vector trọng của nơ-ron chiến thắng được cập nhật mà các vector trọng của các nơ-ron lân cận (hay gọi là "láng giềng") với nơ-ron chiến thắng cũng được cập nhật[21,22].

Để xác định vùng lân cận của nơ-ron chiến thắng hay gọi là vùng chiến thắng, hàm lân cận Gaussian được áp dụng. Hàm lân cận Gaussian được trình bày bởi công thức:

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

Trong đó, $r$là kho ảng cách từ nơ-ron lân cận đến nơ-ron chiến thắng.

$$r = \sqrt{(i_0 - i)^2 + (j_0 - j)^2} \quad (2)$$

Với i_0, j_0 là chỉ số dòng, cột của nơ-ron chiến thắng trên lớp ra. Và $\sigma(t)$ là hàm được sử dụng cho việc xác định không gian lân cận nơ-ron chiến thắng với số lần lặp, giá trị của σ giảm dần[21].

$$\sigma(t) = \sigma_0 e^{-\frac{t}{\tau_1}} \quad (3)$$

Trong đó, ($\tau_1$là hằng số, $\sigma_0 = \sqrt{m}$, t là số lần lặp).

Trong đó, lớp vào là các vector nhập có kích thước n, lớp ra: gồm các node (các nơ-ron) được bố trí trên một lưới (bản đồ). Mỗi nơ-ron có vị trí xác định, tại mỗi nơ-ron lưu giữ một vector trọng số (weight vector) có số chiều bằng với số chiều của vector nhập.

Thử nghiệm phương pháp khám phá cộng đồng

Dữ liệu dùng thử nghiệm phương pháp khám phá cộng đồng là kết quả tìm được từ mô hình TART[24]. Dưới đây trình bày một số mẫu vector nhập trong **Bảng 1.**

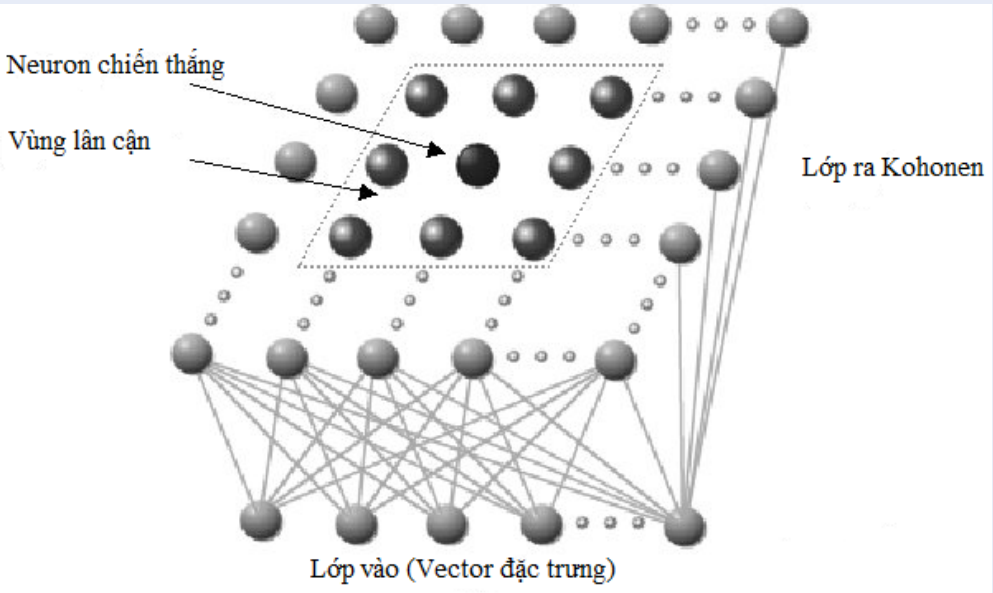
Mô tả dữ liệu thử nghiệm

Bảng 1 là tập 10 vector chủ đề quan tâm đến 6 chủ đề (từ T-0 đến T-6) của 10 cá nhân tham gia trao đổi trong giai đoạn tháng 01-2015. Như vậy, xét trên từng vector, mỗi vector có 7 thành phần. Từng thành phần đó là mức độ quan tâm đến từng chủ đề của cá nhân. Cụ thể, mẫu dữ liệu tại **Bảng 1** là mẫu các vector chủ đề quan tâm của cá nhân trên MXH là kết quả của mô hình TART[14].

Thử nghiệm phương pháp khám phá và trực quan hoá cộng đồng

Gọi C_i là một cụm trên lớp ra Kohonen, C_i được tạo bằng cách tính khoảng cách từ vector nhập đến vector trong tương ứng với cụm đó và gán vector nhập vào cụm có khoảng cách nhỏ nhất bằng phương pháp mạng nơ-ron Kohonen. Kết quả là tại mỗi nơ-ron trên lớp ra Kohonen tương ứng với một tập các đối tượng có chứa các thuộc tính (số cá nhân, tập chủ đề quan tâm) thuộc từng nơ-ron tương ứng là từng cụm (cộng đồng).

- Kích thước lớp ra Kohonen : 14 x 14 (196 nơ-ron).



Hình 6: **Xác định nơ-ron chiến thắng và vùng lân cận của mạng nơ-ron Kohonen [22].**

Bảng 1: **Một số mẫu vector quan tâm chủ đề (vector nhập) của cá nhân tại tháng 01-2015**

Vector	Chủ đề							
	T-0	**T-1**	**T-2**	**T-3**	**T-4**	**T-5**	**T-6**	**Cá nhân**
$\vec{v}_1$	0,47922	0,0	0,43396	0,60427	0,44592	0,3247	0,0	Tuan
$\vec{v}_2$	0,38182	0,36	0,72414	0,72703	0,34163	0,0	0,0	Minh Nguyễn
$\vec{v}_3$	0,33333	0,32075	0,46642	0,35593	0,33333	0,35712	0,41772	Thùy Dương
$\vec{v}_4$	0,61194	0,0	0,56522	0,0	0,31646	0,0	0,0	Ánh Trăng
$\vec{v}_5$	0,40241	0,50124	0,43301	0,0	0,34608	0,3428	0,31343	Hà Nguyễn
$\vec{v}_6$	0,33333	0,74787	0,36456	0,0	0,36232	0,0	0,0	alibaba
$\vec{v}_7$	0,63971	0,35199	0,54959	0,47916	0,44037	0,38475	0,49136	huynd1995
$\vec{v}_8$	0,56479	0,44286	0,65217	0,34884	0,30612	0,3717	0,0	Trung
$\vec{v}_9$	0,7712	0,64083	0,42059	0,50435	0,39593	0,34884	0,34226	Hung
$\vec{v}_10$	0,72819	0,33635	0,43336	0,50981	0,3573	0,45018	0,43044	Nguyên Đức

- Mỗi vector nhập có 25 thành phần tương ứng 25 chủ đề.
- Thời gian : tháng 01-2015
- Số cá nhân tham gia trong tháng 01-2015: 7444
- Kết quả thử nghiệm 1: số cộng đồng khám phá là 60.

Trên **Hình 5,** với từng nơ-ron có màu sậm và nhạt tương ứng với số lượng cá nhân nhiều hay ít tham gia vào cộng đồng. Màu sắc trên mỗi nơ-ron càng đậm đại diện cho số cá nhân trong cộng đồng nhiều hơn những nơ-ron có màu nhạt hơn hoặc cộng đồng không có bất kỳ cá nhân nào (nơ-ron trống không tồn tại cộng đồng).

Mỗi cộng đồng tồn tại 2 thành phần chính là chủ đề quan tâm của cộng đồng và số cá nhân tham gia vào cộng đồng. Chẳng hạn trên **Hình 7,** cộng đồng 35 tại nơ-ron 79 có 14 cá nhân tham gia và cùng quan tâm đến 07 chủ đề (xem danh sách các chủ đề cộng đồng 35 quan tâm được trình bày tại **Hình 8**).

Hình 9 trình bày trực quan kết quả khám phá cộng đồng bao gồm các đặc trưng như cá nhân tham gia và chủ đề quan tâm của của cộng đồng. **Hình 10** trình bày kết quả khám phá cộng đồng quan tâm đến chủ đề 5 trong giai đoạn tháng 01-2015.

Quan sát trong **Bảng 2** nhận thấy rằng, 19 cộng đồng được chọn ngẫu nhiêu trong 41 cộng đồng (xem **Hình 7**) quan tâm đến 15 chủ đề.

Trên **Hình 11**, mỗi cộng đồng thể hiện rõ được xác suất quan tâm đến từng chủ đề cụ thể. Chẳng hạn, cộng đồng 1 quan tâm đến chủ đề T1 là 0,01595. Đây là chủ đề có xác suất quan tâm cao nhất trong khoảng thời gian tháng 01-2015 của cộng đồng số 1.

Trong **Hình 12**, mỗi cộng đồng thể hiện rõ được số lượng cá nhân tham gia. Chẳng hạn, tham gia cộng đồng 14 có 659 cá nhân chiếm 9% và cộng đồng 7 có số cá nhân tham gia cao nhất là 698 chiếm 9% trên tổng số cá nhân tham gia tất cả cộng đồng trong khoảng thời gian tháng 01-2015.

Khảo sát sự biến thiên số cộng đồng dựa trên lớp ra Kohonen

Sự biến thiên số cá nhân tham gia cộng đồng *c* được biết dựa trên tần suất thay đổi số cá nhân *a* trong cộng đồng. Ký hiệu là A (c, t, N_a). Trong đó $c \in C$ là cộng đồng, t là thời gian và N_a là số cá nhân tham gia trong cộng đồng c (hay nói cách khác N_a là số cá nhân trong cộng đồng c) theo từng khoảng thời gian t.

Mỗi cộng đồng có nhiều cá nhân trong từng giai đoạn thời gian. Tuy nhiên, cá nhân trong cộng đồng cũng là đặc trưng cho cộng đồng đó và việc xác định sự thay đổi số cá nhân trong cộng đồng cũng dựa vào cơ sở này. Sự thay đổi này thể hiện qua sự khác nhau giữa thành phần của hai tập số cá nhân trong cộng đồng tại thời điểm $t-1$ là A (c, t -1,N_a) và tại thời điểm t là A (c, t,N_a) mà số cá nhân tham gia cộng đồng. Để đo lường mức độ thay đổi (tính động) số cá nhân a của cộng đồng c tại thời điểm t, bài nghiên cứu xây dựng độ đo $\partial_\theta(c, t-1, t, N_a)$. Độ đo này là tỉ lệ giữa: hiệu số (số cá nhân N_a và phần giao giữa số cá nhân trong cộng đồng tại thời điểm t-1 với cá nhân trong cộng đồng tại thời điểm t) chia cho cá nhân để N_a, giá trị của $\partial_\theta(c, t-1, t, N_a)$ nằm trong khoảng [0,1]:

- Nếu giá trị đạt ở ngưỡng 1 thì tập N_a thường xuyên được thay đổi bởi cộng đồng c
- Ngược lại nếu giá trị đạt ngưỡng 0 nghĩa là số cá nhân trong cộng đồng không thay đổi trong khoảng thời gian từ t -1 đến t. Giá trị ∂_θ được tính bởi công thức (4):

$$\partial_\theta(c, t-1, t, N_a) = \frac{N_a - |A(c,t-1,N_a) \cap A(c,t,N_a)}{N_a} \in [0,1] \quad (4)$$

Từng giai đoạn thời gian, số lượng cá nhân cũng như số cộng đồng tham gia trên MXH cũng có sự thay đổi. Dựa trên lớp ra Kohonen, bài nghiên cứu khảo sát sự biến số cộng đồng tham gia. **Hình 13** trình bày kết quả phân tích sự biến thiên các đặc trưng trong cộng đồng và số cộng đồng tham gia MXH quan tâm trao đổi 15 chủ đề trong trường đại học theo từng giai đoạn thời gian năm 2015.

Dựa trên kết quả trên **Hình 13**, **Hình 14** thể hiện kết quả phân tích sự biến thiên số cộng đồng trên dữ liệu Facebook và 15 chủ đề quan tâm của cộng đồng trong 12 giai đoạn thuộc năm 2014.

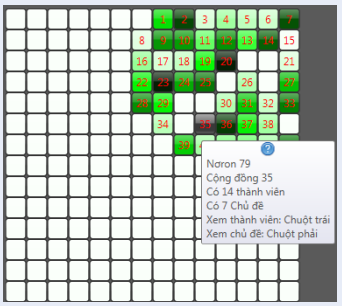
Kết quả thể hiện trên **Hình 12** chỉ ra rằng, trong từng giai đoạn thời gian, số lượng cộng đồng quan tâm đến 15 chủ đề (được khảo sát) có sự thay đổi. Trong đó, tháng 01-2014 có số cộng đồng tham gia nhiều nhất là 62 và tháng 11-2014 có số cộng đồng tham gia ít nhất là 30.

Đánh giá kết quả thử nghiệm phương pháp khám phá cộng đồng và thảo luận

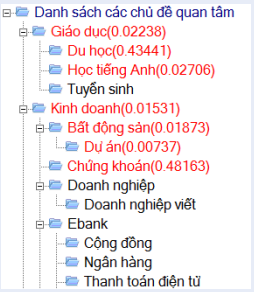
So sánh với phương pháp gom cụm K-Medoids

Bên cạnh việc áp dụng các hệ số Precision, Recall và độ đo F để đánh giá kết quả thử nghiệm, bài nghiên cứu còn áp dụng giá trị RMSSTD[25] (Root Mean Square Standard Deviation) và giá trị RS[26] (R-Squared) để so sánh kết quả giữa phương pháp gom cụm đề xuất trong bài nghiên cứu và giải thuật K-Medoids).

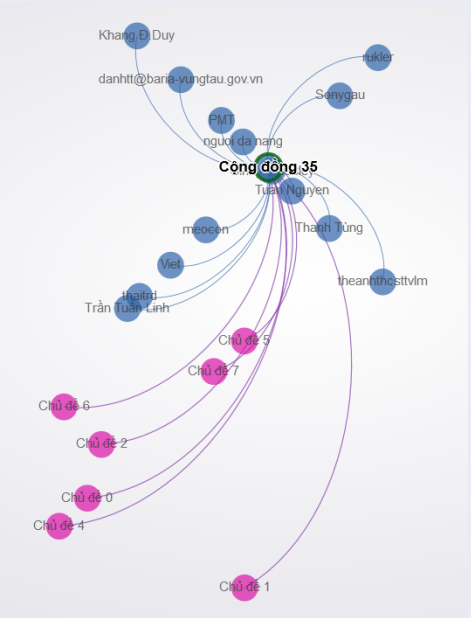
Giá trị RMSSTD là một phương pháp được sử dụng để đo chất lượng của giải thuật gom cụ bằng công thức



Hình 7: **Trực quan hóa kết quả khám phá cộng đồng cá nhân trong tháng 01-2015 hiển thị trực quan trên lớp ra Kohonen.**



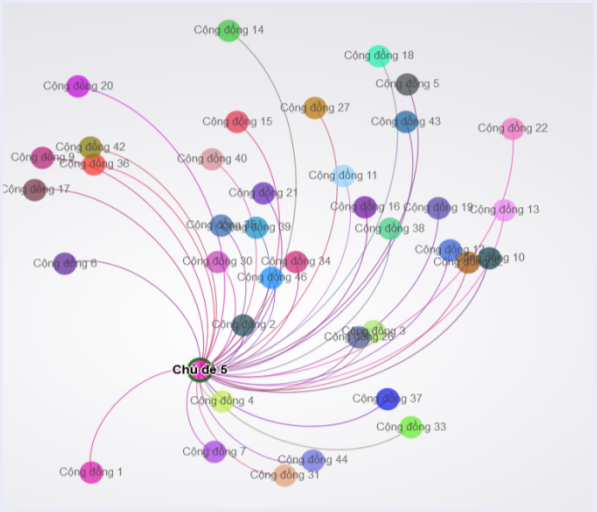
Hình 8: **Danh sách các chủ đề và xác suất quan tâm của cộng đồng 35 trên lớp ra Kohonen.**



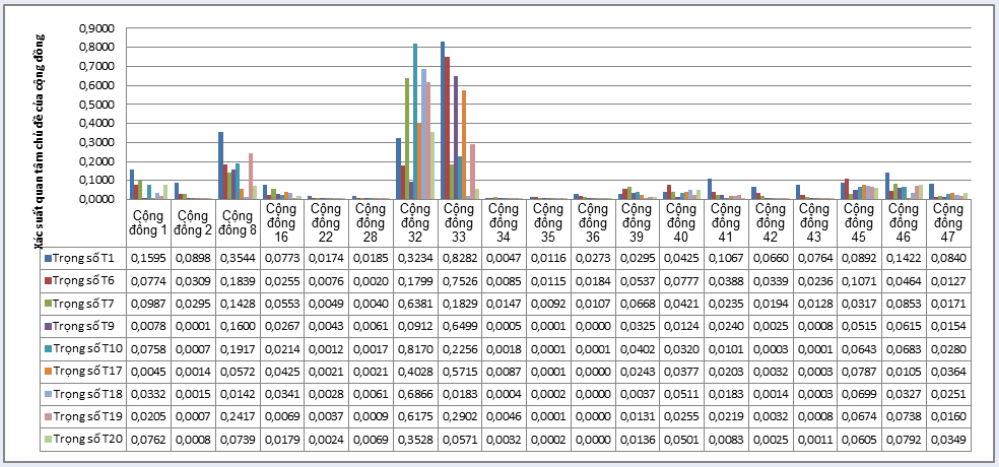
Hình 9: **Trực quan hóa kết quả cộng đồng 35 và các đặc trưng trong cộng đồng.**

Bảng 2: Vector trọng $\overrightarrow{wi}$v với thành phần là xác suất quan tâm chủ đề của từng cộng đồng trong giai đoạn tháng 01-2015

	Trọng số T-1	Trọng số T-6	Trọng số T-7	Trọng số T-9	Trọng số T-10	Trọng số T-17	Trọng số T-18	Trọng số T-19	Trọng số T-20
Cộng đồng 1	0,1595	0,0774	0,0987	0,0078	0,0758	0,0045	0,0332	0,0205	0,0762
Cộng đồng 2	0,0898	0,0309	0,0295	0,0001	0,0007	0,0014	0,0015	0,0007	0,0008
Cộng đồng 8	0,3544	0,1839	0,1428	0,1600	0,1917	0,0572	0,0142	0,2417	0,0739
Cộng đồng 16	0,0773	0,0255	0,0553	0,0267	0,0214	0,0425	0,0341	0,0069	0,0179
Cộng đồng 22	0,0174	0,0076	0,0049	0,0043	0,0012	0,0021	0,0028	0,0037	0,0024
Cộng đồng 28	0,0185	0,0020	0,0040	0,0061	0,0017	0,0021	0,0061	0,0009	0,0069
Cộng đồng 32	0,3234	0,1799	0,6381	0,0912	0,8170	0,4028	0,6866	0,6175	0,3528
Cộng đồng 33	0,8282	0,7526	0,1829	0,6499	0,2256	0,5715	0,0183	0,2902	0,0571
Cộng đồng 34	0,0047	0,0085	0,0147	0,0005	0,0018	0,0087	0,0004	0,0046	0,0032
Cộng đồng 35	0,0116	0,0115	0,0092	0,0001	0,0001	0,0001	0,0002	0,0001	0,0002
Cộng đồng 36	0,0273	0,0184	0,0107	0,0000	0,0001	0,0000	0,0000	0,0000	0,0000
Cộng đồng 39	0,0295	0,0537	0,0668	0,0325	0,0402	0,0243	0,0037	0,0131	0,0136
Cộng đồng 40	0,0425	0,0777	0,0421	0,0124	0,0320	0,0377	0,0511	0,0255	0,0501
Cộng đồng 41	0,1067	0,0388	0,0235	0,0240	0,0101	0,0203	0,0183	0,0219	0,0083
Cộng đồng 42	0,0660	0,0339	0,0194	0,0025	0,0003	0,0032	0,0014	0,0032	0,0025
Cộng đồng 43	0,0764	0,0236	0,0128	0,0008	0,0001	0,0003	0,0003	0,0008	0,0011
Cộng đồng 45	0,0892	0,1071	0,0317	0,0515	0,0643	0,0787	0,0699	0,0674	0,0605
Cộng đồng 46	0,1422	0,0464	0,0853	0,0615	0,0683	0,0105	0,0327	0,0738	0,0792
Cộng đồng 47	0,0840	0,0127	0,0171	0,0154	0,0280	0,0364	0,0251	0,0160	0,0349



Hình 10: **Trực quan hóa kết quả khám phá chủ đề 5 được các cộng đồng quan tâm.**



	Cộng đồng 1	Cộng đồng 2	Cộng đồng 8	Cộng đồng 16	Cộng đồng 22	Cộng đồng 28	Cộng đồng 32	Cộng đồng 33	Cộng đồng 34	Cộng đồng 35	Cộng đồng 36	Cộng đồng 39	Cộng đồng 40	Cộng đồng 41	Cộng đồng 42	Cộng đồng 43	Cộng đồng 45	Cộng đồng 46	Cộng đồng 47
Trọng số T1	0,1595	0,0898	0,3544	0,0773	0,0174	0,0185	0,3234	0,8282	0,0047	0,0116	0,0273	0,0295	0,0425	0,1067	0,0660	0,0764	0,0892	0,1422	0,0840
Trọng số T6	0,0774	0,0309	0,1839	0,0255	0,0076	0,0020	0,1799	0,7526	0,0085	0,0115	0,0184	0,0537	0,0777	0,0388	0,0339	0,0236	0,1071	0,0464	0,0127
Trọng số T7	0,0987	0,0295	0,1428	0,0553	0,0049	0,0040	0,6381	0,1829	0,0147	0,0092	0,0107	0,0668	0,0421	0,0235	0,0194	0,0128	0,0317	0,0853	0,0171
Trọng số T9	0,0078	0,0001	0,1600	0,0267	0,0043	0,0061	0,0912	0,6499	0,0005	0,0001	0,0000	0,0325	0,0124	0,0240	0,0025	0,0008	0,0515	0,0615	0,0154
Trọng số T10	0,0758	0,0007	0,1917	0,0214	0,0012	0,0017	0,8170	0,2256	0,0018	0,0001	0,0001	0,0402	0,0320	0,0101	0,0003	0,0001	0,0643	0,0683	0,0280
Trọng số T17	0,0045	0,0014	0,0572	0,0425	0,0021	0,0021	0,4028	0,5715	0,0087	0,0001	0,0000	0,0243	0,0377	0,0203	0,0032	0,0003	0,0787	0,0105	0,0364
Trọng số T18	0,0332	0,0015	0,0142	0,0341	0,0028	0,0061	0,6866	0,0183	0,0004	0,0002	0,0000	0,0037	0,0511	0,0183	0,0014	0,0003	0,0699	0,0327	0,0251
Trọng số T19	0,0205	0,0007	0,2417	0,0069	0,0037	0,0009	0,6175	0,2902	0,0046	0,0001	0,0000	0,0131	0,0255	0,0219	0,0032	0,0008	0,0674	0,0738	0,0160
Trọng số T20	0,0762	0,0008	0,0739	0,0179	0,0024	0,0069	0,3528	0,0571	0,0032	0,0002	0,0000	0,0136	0,0501	0,0083	0,0025	0,0011	0,0605	0,0792	0,0349

Hình 11: **Kết quả phân tích cộn g đồng dựa trên bảng kết quả 2.**

(5), nếu giá trị của RMSSTD thấp hơn thì kết quả gom cụm tốt hơn.

$$RMSSTD = \frac{\sqrt{\sum_{j=1..p}^{i=1..k} \sum_{a=1}^{n_{ij}} (x_a - \bar{x}_{ij})^2}}{\sum_{j=1..p}^{i=1..k} (n_{ij} - 1)} \quad (5)$$

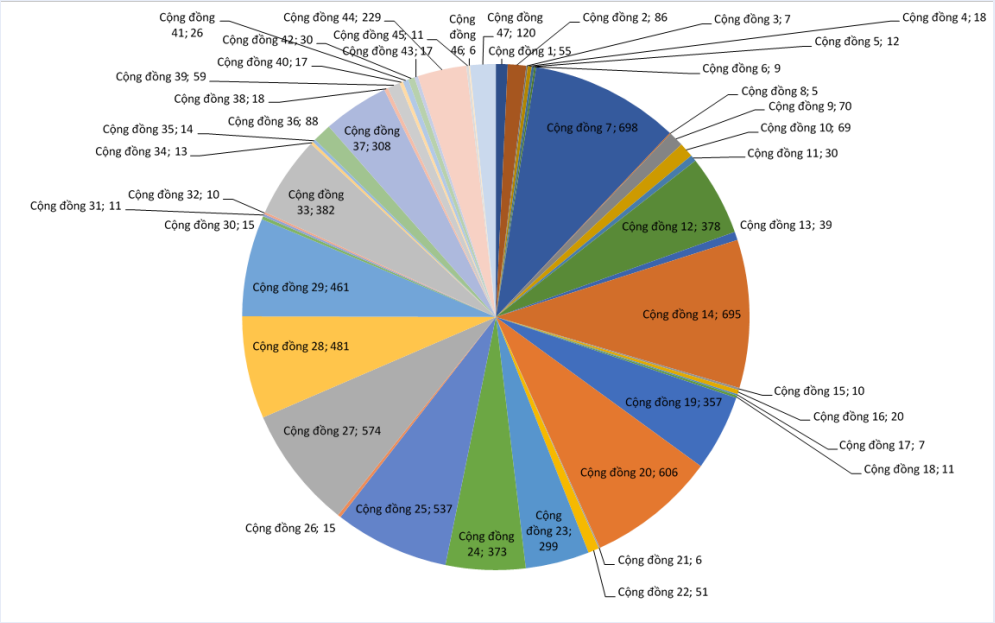
Trong đó k là số cụm, p là số biến độc lập trong tập dữ liệu, $\bar{x}_{ij}$ là trung bình dữ liệu của biến j và cụm i, n_{ij} là số dữ liệu trong biến p và cụm k.

Với những giá trị RMSSTD, trung bình của RMSSTD được tính dựa trên 1000 giao tác cho mỗi lần bộ dữ liệu được thực hiện. Công thức (6) tính trung bình giá trị RMSSTD:

$$Trung\ bình\ RMSSTD = \frac{(Tổng\ giá\ trị\ của\ RMSSTD\ từ\ 1000\ giao\ tác\ mà\ bộ\ dữ\ liệu\ được\ thực\ hiện)}{1000} \quad (6)$$

Giá trị RS được sử dụng để xác định xem có sự khác biệt đáng kể nào giữa các đối tượng dữ liệu trong các cụm khác nhau và giữa các đối tượng dữ liệu trong cùng một nhóm có độ tương tự cao. Nếu giá trị RS bằng 0, thì không có sự khác biệt giữa các cụm. Mặt khác, nếu giá trị RS bằng 1, thì kết quả gom cụm là tối ưu. Giá trị RS được tính bằng công thức (7, 8 và 9):

$$RS = \frac{SS_t - SS_w}{SS_t} \quad (7)$$



Hình 12: **Kết quả phân tích cộng đồng trong tháng 01-2015.**

$$SS_t = \sum_{j=1}^{p} \sum_{a=1}^{n_j} (x_a - \bar{x}_j)^2 \quad (8)$$

$$SS_w = \sum_{j=1}^{i=1..k} \sum_{a=1}^{n_{ij}} (x_a - \bar{x}_{ij})^2 \quad (9)$$

Trong đó SS_t là tổng bình phương khoảng cách giữa tất cả các biến, SS_w là tổng bình phương khoảng cách giữa tất cả đối tượng dữ liệu trong cùng một cụm, trong đó k là số cụm, p là số biến độc lập trong tập dữ liệu, $\bar{x}_{ij}$ là trung bình dữ liệu của biến j và cụm i, n_{ij} là số dữ liệu trong biến p và cụm k.

Giá trị trung bình của RS được tính dựa trên 1000 lần lặp của mỗi lần bộ dữ liệu được thực hiện. Giá trị này được tính bằng công thức (10).

$$\textit{Trung bình RS} = \frac{\textit{(Tổng giá trị của RS từ 1000 lần lặp lại tập dữ liệu)}}{1000} \quad (10)$$

Kết quả thử nghiệm và thảo luận

Thử nghiệm phương pháp đánh giá, các bộ dữ liệu là tập vector từ kết quả của mô hình TART (**Bảng 1**) được bài nghiên cứu sử dụng cho việc thử nghiệm các phương pháp gom cụm để tìm ra giá trị trung bình của RMSSTD và RS. Kết quả thử nghiệm này được lặp lại 1000 lần để cung cấp kết quả ổn định và đáng tin cậy và số lượng các cụm k cũng được thay đổi để có thêm điều kiện so sánh các phương pháp và giải thuật.

Trong **Bảng 3,** cho thấy các giá trị trung bình RMSSTD, phương pháp mạng nơ-ron Kohone cho kết quả RMSSTD thấp nhất cho tất cả các lựa chọn số cụm. Điều này cho thấy rằng, phương pháp mạng nơ-ron Kohonen có kết quả thực hiện vượt trội hơn so với giải thuật K-Medoids.

Trong thử nghiệm này, hai kỹ thuật gom cụm được so sánh dựa trên giá trị RMSSTD và RS (**Bảng 4**) cho bộ dữ liệu thực tế từ kết quả mô hình chủ đề TART. Kết quả cho thấy rằng thuật toán phương pháp mạng nơ-ron Kohonen (SOM) mang lại những giá trị RMSSTD là thấp nhất và giá trị RS là cao nhất.

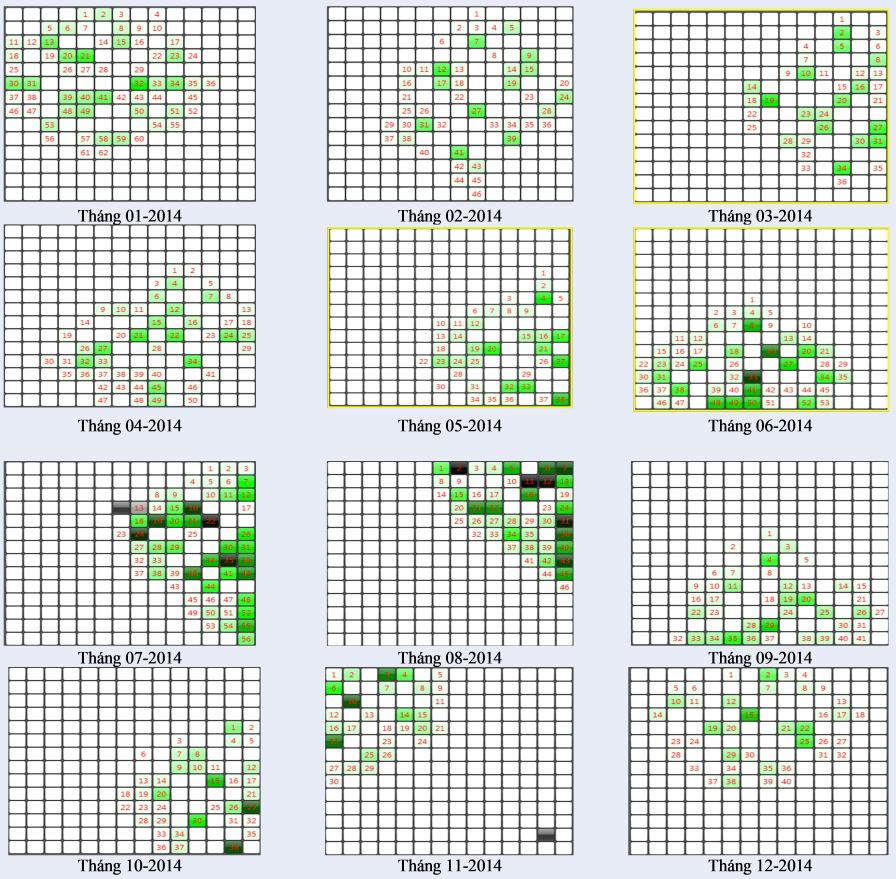
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Bài nghiên cứu đã giải quyết được hai vấn đề quan trọng đóng góp về mặt khoa học và thực tiễn trong lĩnh vực khám phá cộng đồng:

Thứ nhất là xây dựng phương pháp khám phá cộng đồng cá nhân dựa theo mô hình chủ đề có yếu tố thời gian và phân tích sự biến thiên đặc trưng của cộng đồng.

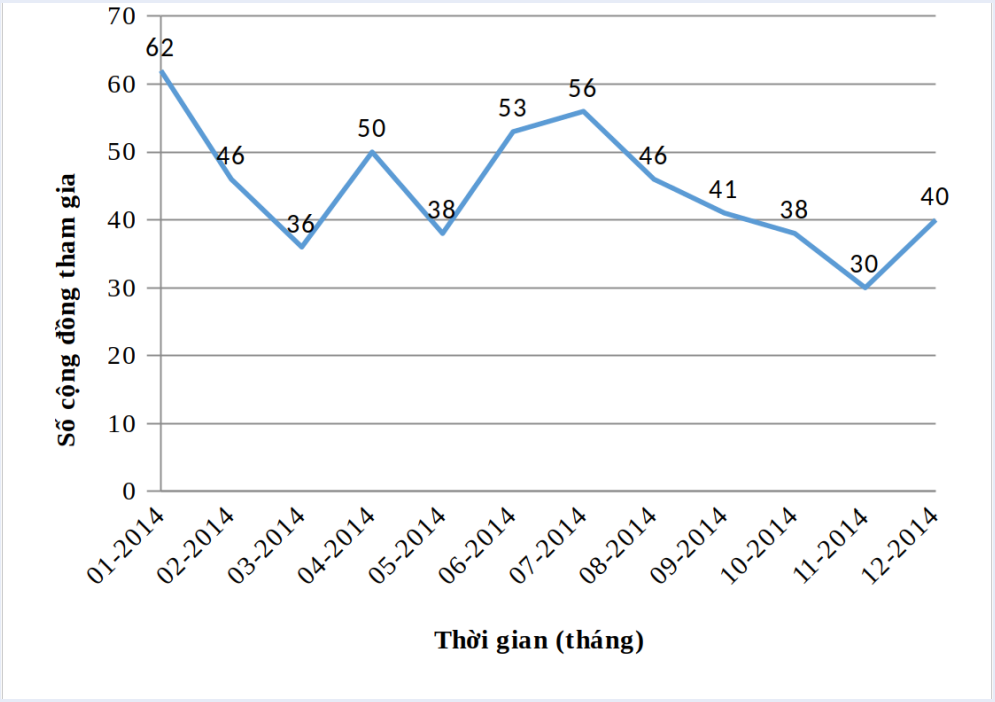
Phương pháp này giúp tìm ra nhóm cá nhân có cùng chủ đề và mức độ quan tâm chủ đề trong từng giai đoạn thời gian. Áp dụng phương pháp huấn luyện mạng nơ-ron Kohonen để khám phá cộng đồng những cá nhân cùng quan tâm đến từng chủ đề cụ thể được gọi là cộng đồng cá nhân theo chủ đề dựa trên tập vector đầu ra của mô hình TART. Trong đó,



Hình 13: **Số cộng đồng tham gia MXH quan tâm đến 15 chủ đề theo từng giai đoạn thời gian thể hiện trên lớp ra Kohonen.**

Bảng 3: **Bảng kết quả giá trị trung bình RMSSTD dựa trên thử nghiệm hai phương pháp gom cụm**

Số cụm k	Kohonen	K-Medoids
2	**0,69635**	0,75288
3	**0,58297**	0,65064
4	**0,52873**	0,59444
5	**0,49807**	0,55666
6	**0,47517**	0,52774
7	**0,45634**	0,50502
8	**0,44195**	0,48648



Hình 14: **Sự biến thiên số cộng đồng theo từng giai đoạn thời gian.**

Bảng 4: **Bảng kết quả giá trị trung bình RS dựa trên thử nghiệm hai phương pháp gom cụm**

Số cụm k	Kohonen	K-Medoids
2	**0,49659**	0,40112
3	**0,63921**	0,55356
4	**0,70391**	0,63431
5	**0,74951**	0,68794
6	**0,78086**	0,72456
7	**0,8034**	0,75273
8	**0,82022**	0,77574

phương pháp khám phá cộng đồng tính được phân bố chủ đề theo từng cộng đồng và tính cụ thể chủ đề được những cộng đồng nào quan tâm và mức độ quan tâm. Kết quả khám phá cộng đồng được trực quan hoá trên lớp ra Kohonen. Sau đó, dựa vào kết quả trên lớp ra Kohonen, bài báo phân tích sự biến thiên các đặc trưng của cộng đồng như: chủ đề quan tâm và cá nhân tham gia cộng đồng theo từng giai đoạn thời gian.

Thứ hai làđể thực nghiệm các mô hình và phương pháp, nghiên cứu đã thử nghiệm và đánh giá mô hình và phương pháp trên hai tập dữ liệu thông điệp tiếng Việt được thu thập từ MXH trong trường đại học và trang báo điện tử VnExpress.net. Để tiến hành thử nghiệm, nghiên cứu đã xây dựng một hệ thống phần mềm phân tích MXH thực hiện đầy đủ các bước trong phương pháp khám phá cộng đồng. Kết quả thực nghiệm đã cho thấy được hướng ứng dụng nghiên cứu của bài báo và khả năng khai thác hiệu quả của phần mềm vào ứng dụng thực tế.

Hạn chế và hướng phát triển

Kết quả nghiên cứu tập trung vào việc giải quyết các bài toán khám phá cộng đồng cá nhân trên MXH. Trong nghiên cứu tiếp theo, chúng tôi sẽ tập trung phân tích ảnh hưởng lan truyền chủ đề của cộng đồng trên MXH. Mục tiêu phân tích ảnh hưởng lan truyền thông điệp trên MXH nhằm xác định " đường đi "

và tìm ra nguồn gốc thông tin. Xây dựng hệ thống khoảng thời gian (có tính chất overlap) để phân tích trực tuyến MXH theo nhiều khoảng thời gian khác nhau.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Kinh tế - Luật, ĐHQG-HCM thông qua đề tài với mã số CS /2018-01 và Phòng Nghiên cứu Kinh doanh Thông minh (BI-LAB), Khoa Hệ thống Thông tin, Trường Đại học Kinh tế - Luật.

DANH MỤC TỪ VIẾT TẮT

MXH: mạng xã hội
ART: Author-Recipient-Topic
TART: Temporal-Author-Recipient-Topic
SOM: Self-Organizing Map
GT: Group-Topic
CUT: Community-User-Topic
ATC: Author-Topic-Community
RMSSTD: Root Mean Square Standard Deviation
RS: R-Squared

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kì xung đột lợi ích nào trong công bố bài báo.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Tác giả Hồ Trung Thành, Trần Duy Thanh và Nguyễn Quang Hưng cùng đóng góp về ý tưởng, mục tiêu, lựa chọn phương pháp nghiên cứu và các vấn đề liên quan đến trực quan hoá dữ liệu. Tác giả Hồ Trung Thành đã đóng góp đề xuất mô hình phân tích dữ liệu mạng xã hội và phương pháp và thực nghiệm khám phá cộng đồng, đánh giá kết quả thực nghiệm. Tác giả Trần Duy Thanh đã đóng góp về thu thập dữ liệu, xây dựng hệ thống phần mềm phân tích dữ liệu. Tác giả Nguyễn Quang Hưng đóng góp về xử lý dữ liệu đầu vào, khảo sát sự biến thiên của cộng đồng và đánh giá kết quả thực nghiệm.

TÀI LIỆU THAM KHẢO

1. Durgesh MS, Moiz M. Sentiment Analysis on Social Networking: A Literature Review. International Journal on IJRITCC. 2015;3(2):022–027.
2. Aggarwal C. Social Network Data Analytics. IBM Thomas J. Watson Research Center; 2011.
3. Kirchhoff L. Applying Social Network Analysis to Information Retrieval on the World Wide Web: A Case Study of Academic Publication Space. Switzerland: The University of St. Gallen; 2010.
4. Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press; 1994.
5. Abdelbary HA, Abeer ME, Reem BT. Utilizing Deep Learning for Content-based Community Detection. In: Science and Information Conference, UK. IEEE; 2014. p. 777–784.
6. Aggarwal C, Subbian K. Event detection in social streams. In: Proceedings of the 2012 SIAM international conference on data mining; 2012. p. 624–635.
7. Li C, Cheung WK, Ye Y, Zhang X, Chu D, Li X. The Author-Topic-Community model for author interest profiling and community discovery. London: Springer-Verlag; 2014. p. 74–85.
8. Zhou D, Manavoglu E, Li J, Giles CL, Zha H. Probabilistic models for discovering e-communities. WWW '06: Proceedings of the 15th international conference on World Wide Web, ACM. 2006;p. 173–182.
9. Pathak P, Delong C, Banerjee A, Erickson K. Social topic models for community extraction. In: The 2nd SNA-KDD Workshop. vol. 8; 2008.
10. Wang X, Mohanty N, Mccallum A. Group and topic discovery from relations and their attributes. Advances in Neural Information Processing Systems. 2006;18:1449–1456.
11. Adham B, Ognjen A, Dinh P, Svetha V. Discovering Topic Structures of a Temporally Evolving Document Corpus. Journal: Knowledge and Information Systems. 2015;arXiv:1512.08008v1:1–53.
12. Zhou D, Councill I, Zha H, Lee GC. Discovering Temporal Communities from Social Network Documents. IEEE ICDM. 2007;p. 745–750.
13. Freeman LC. Visualizing Social Networks. Journal of Social Structure. 2000;Available from: http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.
14. Yin Z, Cao L, Gu Q, Han J. Latent community Topic Analysis: Integration of Community Discovery with Topic Modeling. ACM Transactions on Intelligent Systems and Technology. 2012;3(4):1–21.
15. Alexandru B, Markus D, Nicolai R. Content and communication based sub-community detection using probabilistic topic models. IADIS International Conference Intelligent Systems and Agents © IADIS. 2009;.
16. Fani H, Zarrinkalam F, Zhao X. Temporal Identification of Latent Communities on Twitter. In: The 9th ACM International Conference on Web Search and Data Mining (WSDM2016). vol. arXiv:1509.04227v1 [cs.SI]; 2016.
17. Rosen-Zvi M, Griffths T, et al. Probabilistic Author-Topic Models for Information Discovery. In: 10th ACM SigKDD, Seattle; 2004. p. 306–315.
18. Yang T, Chi Y, Zhu S, Gong Y, R J. Detecting communities and their evolutions in dynamic social networks-a Bayesian approach. Mach Learn. 2011;82:157–189.
19. Griffiths T. Gibbs Sampling in the generative model of Latent Dirichlet Allocation. 2004;Gruffydd@psych.stanford.edu.
20. Andrew M, Andrés C, Xuerui W. Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research. 2007;30(1):249–272.
21. Kohonen T. Self-Organized Formation of Topologically Correct Feature Maps. Biol Cybern. 1982;43:59–69.
22. Kohonen T. Self-Organization and Associative Memory. Berlin: Springer; 1984.
23. Haykin S. Neural Networks. A Comprehensive Foundation. New Jersey: Prentice-Hall, Inc.; 1999. p. 443–465.
24. Ho T, Do P. Social Network Analysis Based on Topic Model with Temporal Factor. International Journal of Knowledge and Systems Science (IJKSS). 2018;9(1).
25. Halkidi M, Batistakis Y, Vazirgiannis M. Cluster validity methods: Part I. SIGMOD REC. 2002;31(2):40–45.
26. Halkidi M, Batistakis Y, Vazirgiannis M. Clustering validity checking methods: Part II. SIGMOD REC. 2002;31(3):19–27.

Science & Technology Development Journal – Economics - Law and Management, 3(3):311- 326

Open Access Full Text Article

Research Article

Applying topic model combined with Kohonen networks to discover and visualize communities on social networks

Ho Trung Thanh*, Nguyen Quang Hung, Tran Duy Thanh

ABSTRACT

Users are members of communities on social networks. Users' interested topics keep changing, resulting in the change of their communities' interested topics as well. Level, period of time, and interested topics represent features of a community which (i) change upon preferences of each user on social networks for making friends or being interested in topics (based on message content); (ii) are formed or change from online groups of friends or the suggestions to make friends. Hence, the link of users in communities can be viewed as a network of users by their features in social network communities. In this paper, the author studies and proposes a new model for discovering communities using Temporal-Author-Recipient-Topic (TART) model combined with Kohonen neural networks to discover communities of users with the same interested topics over different periods of time. The research goal is achieved through testing models on two Vietnamese datasets (collected from social networks at universities and online newspapers).

Key words: Discovering communities, social network analysis, TART model, Kohonen neural networks, topic model

University of Economics & Law, VNUHCM, Vietnam

Correspondence

Ho Trung Thanh, University of Economics & Law, VNUHCM, Vietnam

Email: thanhht@uel.edu.vn

History

- Received: 19/2/2019
- Accepted: 25/4/2019
- Published: 30/9/2019

DOI : 10.32508/stdjelm.v3i3.572

Check for updates

VNU-HCM Press

Cite this article : Trung Thanh H, Quang Hung N, Duy Thanh T. **Applying topic model combined with Kohonen networks to discover and visualize communities on social networks**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 3(3):311-326.