J Inf Process Syst, Vol.17, No.1, pp.163~177, February 2021
https://doi.org/10.3745/JIPS.04.0206

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Discovering Community Interests Approach to Topic Model with Time Factor and Clustering Methods

Thanh Ho*,** and Tran Duy Thanh*,**

Abstract

Many methods of discovering social networking communities or clustering of features are based on the network structure or the content network. This paper proposes a community discovery method based on topic models using a time factor and an unsupervised clustering method. Online community discovery enables organizations and businesses to thoroughly understand the trend in users' interests in their products and services. In addition, an insight into customer experience on social networks is a tremendous competitive advantage in this era of e-commerce and Internet development. The objective of this work is to find clusters (communities) such that each cluster's nodes contain topics and individuals having similarities in the attribute space. In terms of social media analytics, the method seeks communities whose members have similar features. The method is experimented with and evaluated using a Vietnamese corpus of comments and messages collected on social networks and e-commerce sites in various sectors from 2016 to 2019. The experimental results demonstrate the effectiveness of the proposed method over other methods.

Keywords

Clustering Method, Community Interests, Feature Vectors Social Network, Topic Model, Time Factor, User Experience

1. Introduction

The emergence of online social networks over past decades has resulted in huge increases in personal data and information, human activities, connections and relationships among users or groups, and discussions on their opinions and thoughts [1]. The integration of social relationships among users can improve the accuracy of recommendation results since user preferences are similar or influenced by their connected friends [2]. The large volume of this information can be related to individuals or groups and can be interpreted as nodes in a graph [1]. Analyzing the behaviors of individuals or groups on social networks finds related labels such as those on demographics (e.g., age, gender, and location); labels representing political opinion or religious belief; and many other characteristics capturing aspects of users' information and their behaviors on social networks [1]. These labels often appear in personal data on social networks or are associated with other data objects in the network, such as comments, images, and multimedia data. The discovery of an interest-based community is a way to analyze social networks to find groups of users with social connections in the network and topics of interest [3-7].

Manuscript received May 11, 2020; first revision July 20, 2020; second revision September 24, 2020; third revision October 21, 2020; accepted November 8, 2020.

Corresponding Author: Thanh Ho (hotrungthanh@gmail.com)
* Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam (thanhht@uel.edu.vn, thanhtd@uel.edu.vn)
** Vietnam National University, Ho Chi Minh City, Vietnam

Moreover, labels can help us understand users' interests through their interest in the social networking community for a particular topic. Community plays a vital role in shaping a social network. Community discovery assists in gaining knowledge of customer interests through the products and services offered. Changes in the community are frequently related to the characteristics of the community, such as topics of interest, the number of users, and the degree of interest in the topic at different times. This leads to changes in behavior and in topics of interest among users in the community. As users' interests in topics change over time, corresponding changes in social networking communities occur.

The online community can change for two reasons: (i) acquaintances become friends through other friends or referrals and (ii) users' interests change in discussions in comments and messages on a social network. Therefore, the relationship of online communities is a social network with a combination of users. This relationship is depicted through social networks [4,8]. Owing to each user's properties on social networks, different message content exists in the form of text, images, and multimedia. For a while, a given online community can discuss many topics, and other communities may also discuss these topics.

The objective of this research article is to focus on answering the following questions. How does user experience in communities develop through the content of messages and commentary on social networks? With a specific topic or group, which communities on social networks are interested in exchanging information? What is the variety of exciting topics and user participation currently in the community? Finding answers to these questions is not easy, but the results of this research can help analyze and discover topics of interest and find influential users in the community. These users may aid in developing strategies such as user management in the community of a company, organization, or country. These results can help to understand users in order to implement effective marketing strategies, developing online training in education, and other fields of application.

2. Related Works

Researchers proposed several models to explore groups or individual communities on social networks who are interested in the topic regarding the method that discovers the individual community on social networks in the previous studies. These models were experimented on and reported in scientific articles and emails' content in the English language. The researchers focus on exploring the groups or user communities on social networks that are interested in the same topic [5,6,9-13]. In addition, the articles investigated the social networking community [8,14-18] and Tweet clustering [19] based on the topic model. Some typical models, such as the group-topic (GT) model [10], are built based on the Bayesian network method. The objective of the GT model is to discover hiding users on social networks using users' discussed content analysis. This model is considered to group individuals by topic based on the attributes and content of each individual's discussion on the social network. Applying the topic model with additional elements is grouped by the unsupervised learning method. The GT model considers each individual having a relationship with other individuals online if those individuals have the same behavior and connect message content in the same event. However, this study did not specify community members such as sender and receiver.

The community-user-topic (CUT) model [5] based on the Bayesian network method, the Gibbs sampling technique, and the community discovery method are employed to find the set of users interested in specific topics and formed the communities. However, like some of the other models, the CUT model

[5] ignored the time factor of the discussed topic and users' roles. It is essential to analyze the trend in topics of interest to the user role. The author-topic-community (ATC) model [4] was proposed and published by the authors in 2015. The ATC model focuses on exploiting the main components of author A, community C, and topic T. In the research [4], the authors did not concentrate on exploiting the time factor and analyzing the variation of topics and users of communities on social networks. Besides, the above studies did not pay attention to analyzing the distribution of topics in the community over time, distributing topics of interest in the community, and the changes in users' interest in each topic. Those studies focused on community discovery based on English message data while we reinvestigate and experiment with the proposed model on the Vietnamese corpus collected from social networks.

To deal with the limitations of previous studies, this article proposes a method of community discovery based on a temporal-author-recipient-topic (TART) with the time factor [20] combining the Kohonen neural network to explore the community over time as well as to visualize the results of community discovery based on the Kohonen output layer. We apply the Kohonen training method in different ways. We cluster users having the same exciting topic but different levels of interest. The striking advantage of this grouping is to solve the criteria of predetermining the number of clusters in the clustering method.

3. Discovering Community Interests on Social Networks

3.1 Definitions

A set of communities on the Internet is denoted by *C,* and an under-consideration community is denoted by *c*. We have $c \in C$ [6].

Definition 1 (Social network community). The social networking community is a collection of users who pursue common interests or goals and interact through specific media but can cross geographical and political boundaries [6,21].

Definition 2 (Social network community by topics). Based on the topic model, a community is a collection of users who are interested in common topics. Each user in the community is characterized by an interested topic vector and has a greater degree of interest in the topic in the community than in other communities. Let *c* be a topic community, $c \in C$, where *C* is a set of community. Community is a segment with characteristics such as cluster, denoted by $C = \{C_1, C_2, C_3, C_4,...,C_K\}$, where K is the number of communities, and each C_i community has an interested topic vector set:

(1) Disjoint: $C_i \cap C_j = \emptyset$ if both communities do not have one or many same interested topics.

(2) Intersecting: $\bigcup_{i=1}^{K} C_i = C$.

This article builds and concentrates on Definition 2 to research and experiment with the proposed method.

3.2 Discovering Community Interests by Topic Models and Clustering Methods

A clustering method (community discovery) identifies data clusters where each cluster is a set of similar data. The similarity of the data is described and determined by a distance function that depends

on the method (usually the Euclidean distance function). The purpose of aggregating data clusters is to identify the data density in large, N-dimensional datasets, thereby understanding the input data structure and identifying data clusters with similar characteristics. There are many clustering techniques such as SVM, K-means, K-Medoids, and the Kohonen neural network (also known as the self-organizing map [SOM]) [22]. This neural network was developed by Kohonen [22] in the 1980s to solve the flattening clustering problem. The Kohonen neural network gathers data clusters without specifying the number of clusters in advance. This correlates with the data cluster in this research, which is a large thematic network community with an extremely large, N-dimensional messages dataset, and it is challenging to predefine the number of clusters and communities**.** This research utilizes the Kohonen neural network to visualize the results of community discovery in the network's 2D output [22].

An important feature of the Kohonen neural network is that it can map N-dimensional input vectors onto a one- or two-dimensional map [22-24]. The adjacent vectors in the input space will be near each other on the output map of the Kohonen neural network. This allows solving the problem that brings the N-dimensional interested topic vector (results of the TART model [20]) into a two-dimensional vector to visualize in the network output layer.

A Kohonen neural network consists of a grid of output nodes and N input nodes. Each link between the input and output of the Kohonen neural network corresponds to a weight. According to the nature of the training algorithm on the neural network, clusters near each other in the network will contain highly similar objects with the same features in the community.

4. Method of Discovering Community Interests

4.1 Proposed Method

The users' community of a social network discovery method has two main tasks, based on the topic model used to explore the community. The first task is developing a method for exploring topics and discovering communities based on the topic model, including a time factor (Fig. 1). Through survey, analysis, and evaluation by community discovery models, the article explains how the Kohonen training method is applied. The second task is combining the training the neural net and standardizing the input dataset (a set of users' interested topic vectors for each period, a result of the TART model). From this result, we implement the user community discovery method. The results are shown on the neurons of the Kohonen output layer.

Using the clustering method, community discovery is based on the characteristic vectors of users in each period. This method is implemented as shown in Fig. 1 and has six modules:

The first three modules detect and label the topics of interest. The results obtained in module 2 are a list of latent topics that have not been labeled, but module 3 classifies the topic corresponding to the label. To accomplish this task, we need to create a topic taxonomy, which is built in the same domain as the survey and analysis of the data content. Building a topic taxonomy creates training data sets for text classification and topic labeling; in combination with support vector machine (SVM) [25], topic taxonomy is used to label latent topics, and the result of module 3 is a set of labeled latent topics that are imported into the TART model.

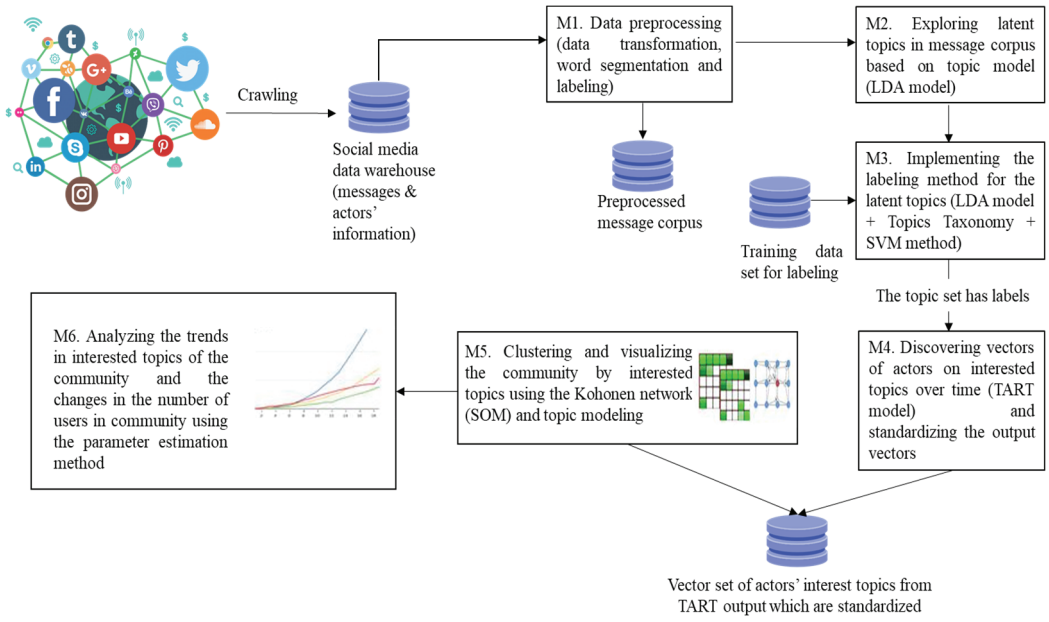


Fig. 1. The interested topic-communities discovery method by the topic model and clustering method.

Module 4. The TART model aims to discover the set of feature vectors (interested topic vectors and input vectors) and then standardize the input vectors. This standardization process provides the data needed for training the Kohonen neural network [23]. Specifically, module 4 standardizes the user topics of interest vectors at different periods according to the TART model results. Then, the input vectors can be used for the neural network training. Because the interested topic vectors of the TART model can give values >1, this does not satisfy the condition that the vector space of the critical vector must be in the range [0, 1].

Module 5. This module discovers and visualizes the community using a Kohonen neural network to gather clusters of users according to interested topics, where each cluster is a community and corresponds to one neuron in the output layer.

Module 6. The typical variation of the community over time is analyzed based on the output layer of the neural network.

4.2 Algorithm for the Proposed Method

The article applies the Kohonen neural network to detect clusters of users according to topics of interest. Based on the set of vectors for users' topics of interest in each period, the training process for clustering is based on the characteristic vectors from the TART model [20]. Each cluster is a community interested in many topics over a period and is found on each neuron in the output layer.

Problems: In the social network $G = <V, E>$, V is a set of users, and E is a set of messages discussed among users. Given a set of users' interested topic vectors, find C communities including users who have the same interested topic and their level of interest over time (see Algorithm 1).

Algorithm 1. Discovering community interests

Given:

Input vectors (vectors of users' interested topics) {v_i} due to the TART model. Vector v_i has m dimensions: $v_i <v_{i1}, v_{i2}, \ldots, v_{im}>$, where m is the number of interested topics.

The components of the input vectors include the set of topics that the user is interested in, the level of interest, and the length of time that the user is interested in the topic.

Find:

Find the list of user communities $C = \{C_1, C_2, C_3, C_4, \ldots, C_K\}$ interested in the topic set from time to time. Characteristics of each community C_i include interested topics, levels of interest, and the number of users participating in the community.

Where K is the number of communities; clusters have the following properties:

- *Disjoint: $C_i \cap C_j = \emptyset$ if the clusters do not have one or many same specific interested topics.*
- *Intersecting: $\bigcup_{i=1}^{K} C_i = C$*

Methods: Using a Kohonen neural network, our method has four steps:

Step 1) Standardizing v_i input vector

Input vector standardization converts vector x to vector x' such that the components of vector x' are standardized with respect to the input vector to Kohonen network training. For example, we have vector $x = (x_1, x_2, x_3, ..., x_n)$ consisting of five elements and needing to be standardized. Vector x is normalized by multiplying x by a positive number c [23].

$$c = \frac{1.0}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}} \quad (1)$$

where x_i is a vector element x. Assuming that we have vector $x = (1.2, 2.3, 3.4, 4.55, 5.6)$, applying formula (1) to standardize the vector yields:

$$c = \frac{1.0}{\sqrt{1.2^2 + 2.3^2 + 3.4^2 + 4.55^2 + 5.6^2}} = 0.1192$$

Then, the value $c = 0.1192$ will be used as a multiplier for vector x. The standardized vector x' is as follows:

$\boldsymbol{x}$**'** = (1.2* 0.1192, 2.3*0.1192, 3.4*0.1192, 4.55*0.1192, 5.6*0.1192)

⇨ $\boldsymbol{x}$**'** = (0.1430, 0.2742, 0.4053, 0.5424, 0.6675)

This is done for all the vectors to form the input vector for the process of community discovery using the Kohonen network training method.

Step 2) Input vector v_i to the Kohonen neural network for the training process.

Step 3) For each $i \in [1,...,n]$ //n is the number of columns and rows of the Kohonen output layer.

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

Find the neuron that has the weight vector w_{ij} closest to the input vector v

Call (i_0, j_0) are the coordinates of the winning neuron. We have $d(v,w_{i0,j0}) = min(d(v,w_{ij}))$, which

is the distance (where $i,j \in [1,...,n]$) and $w_{i0,j0}$ is weight of the winning neuron.

Step 4) Identify neighborhoods and update the winning neurons.

A SOM network applies soft-competition to cluster data. Therefore, not only are the weight vectors of the winning neurons updated, but the neighboring vectors (or "neighbors") with winning neurons are updated as well [22,24] (see Algorithm 2).

Algorithm 2. Finding the winning neurons [22,24]

Given: v, SOM. In which, v is set of input vectors
Find: winning neuron $w_{i_0 j_0}$
Method:
Begin
    Initialization $min = d(v, SOM[0,0])$;
    Initialization $wneuron = SOM[0,0]$;
    Loop $i = 0$ *to* $sqrt(m)$
        Loop $j = 0$ *to* $sqrt(m)$
            If $min > d(v, SOM[i,j])$
            Then
                $min = d(v, SOM[i,j])$;
                $wneuron = SOM[i,j]$;
    Return winning neuron
End

4.3 Experimental Method and Visualization

4.3.1 Experimental data

The dataset used for the experimenting community discovery method is the result obtained from the TART model [20]. Example input vectors are shown in Table 1.

Table 1. Samples for the user-interested topic (T) vectors (feature vectors) in January 2019

Feature vectors	Topics							Users
	T-0	T-1	T-2	T-3	T-4	T-5	T-6	
$\overrightarrow{v_1}$	0.64444	0.34545	0.46826	0	0.33721	0	0	Mr.tajkjd
$\overrightarrow{v_2}$	0.30435	0.44565	0.33333	0.30435	0.33333	0.52941	0	dsvantan
$\overrightarrow{v_3}$	0.39601	0.48718	0	0.35484	0	0.38462	0	nguyen.nhi.334491
$\overrightarrow{v_4}$	0.34694	0.40741	0	0.39227	0	0.36000	0	trang.harry.7
$\overrightarrow{v_5}$	0	0.35135	0	0.41935	0	0.31429	0	anna.vy.334
$\overrightarrow{v_6}$	0	0.36000	0	0.33333	0	0.44828	0.40741	haianh.nguyen.52012
$\overrightarrow{v_7}$	0.48718	0.32431	0	0	0	0	0.31034	quyvan.pham.54
$\overrightarrow{v_8}$	0.40741	0.31034	0	0	0	0.41772	0	su.heo.1656
$\overrightarrow{v_9}$	0.35135	0.33333	0.40741	0	0.30923	0.34545	0	phuc.hanh.9678
$\overrightarrow{v_{10}}$	0.64557	0.90000	0.34884	0.58974	0.33354	0	0.77465	GiámĐốcTàiChính TiềmNăng

There were 921,310 discussed messages, including the content and comments that were collected, and 121,349 user accounts on social networks between 2016 and 2019. This research was concerned with the information about the user's ID when it was issued when joining the website, the user's name, the message, the message sender and receiver, and the time factor. After the training data has been learned via the TART model, the collected results are the set of interested topic vectors for each period.

Table 1 presents a representative set of ten interested topic vectors for six topics (T-0 to T-6) of ten participants in January 2019. Each vector has seven components, each of which has a level of interest for each topic. Specifically, the data sample in Table 1 is the sample of interested topic vectors of users on social networks and a result sample of the TART model.

4.3.2 Experimental methods and visualization

Let C_i be a cluster of the Kohonen output layer, C_i is created by calculating the distance from the input vector to the corresponding weight vector to that cluster. The input vector is then assigned to the cluster with the smallest distance using the Kohonen method. The result is that each neuron in the output layer corresponding to a set of objects with attributes (users, interested topics) belongs to the neuron corresponding to each cluster (community).

- The Kohonen output layer size: 14×14 (196 neurons).
- Each input vector has 15 elements corresponding to 15 topics.
- Time: January 2019.
- The number of users participating in January 2019: 2,244.
- Test result 1: the number of discovered communities is 41.

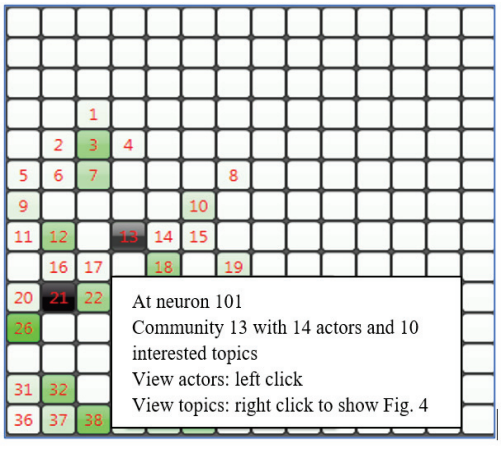


Fig. 2. Visualizing the results of user community discovery in January 2019 on the Kohonen output layer.

Each neuron in the Kohonen output layer in Fig. 2 is shaded according to the number of users participating in the community. Darker neurons indicate that more users participate in the community than lighter neurons. Shading can also indicate that a community does not have any users (the empty neurons mean the community does not exist). Each community contains two traits, the topic of interest and the number of users in the community. For example, in Fig. 2, community 13 at neuron 101 has 61

users who participate in and are interested in ten topics (see list of community 13 with topics presented in Fig. 3). Among these topics, the "Recruitment and Employment - Tuyển dụng việc làm" topic has the highest probability of 0.3401 and is followed by "School security – An ninh học đường" with the probability of 0.04454. "General education – Giáo dục" and "Union activities – Hoạt động đoàn hội" have probabilities of 0.02928 and 0.01975, respectively. The lowest probability of 0.00504 is for the "Study abroad – Du học" topic.

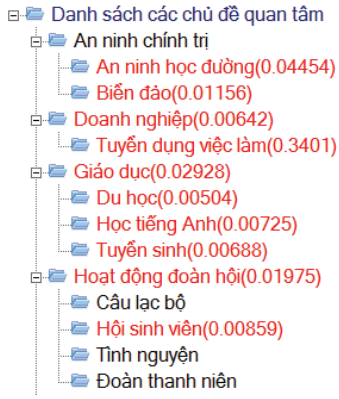


Fig. 3. Part of the topic taxonomy in Vietnamese and probabilities of interested topics in the 13th community of the Kohonen output layer.

Fig. 4 shows the results of community discovery, including characteristics such as user involvement and the interested topic community. Community 13 has several topics in a variety of sectors that interest users. Fig. 5 presents the exploration results of the community with interested topic 11 in January 2019. The "Project – Dự án" topic related to "education" attracts many communities.

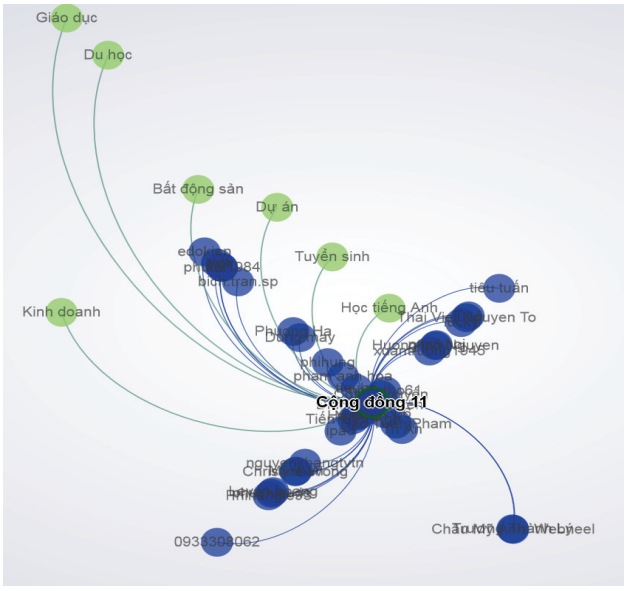


Fig. 4. Visualizing community 11 (cộng đồng 11) results and community features.

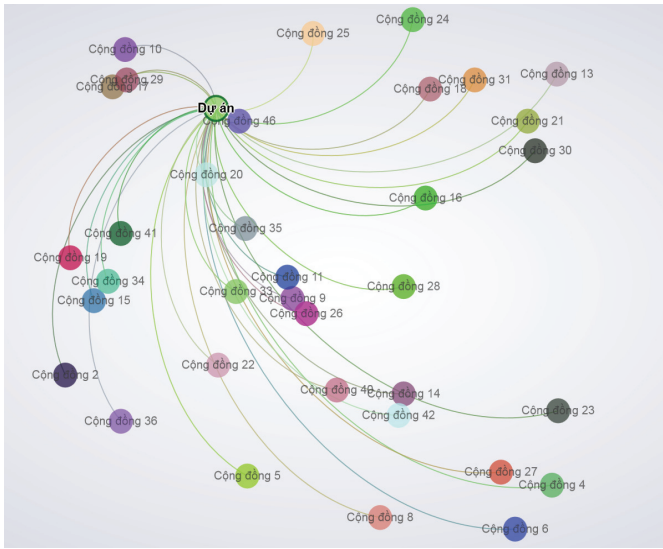
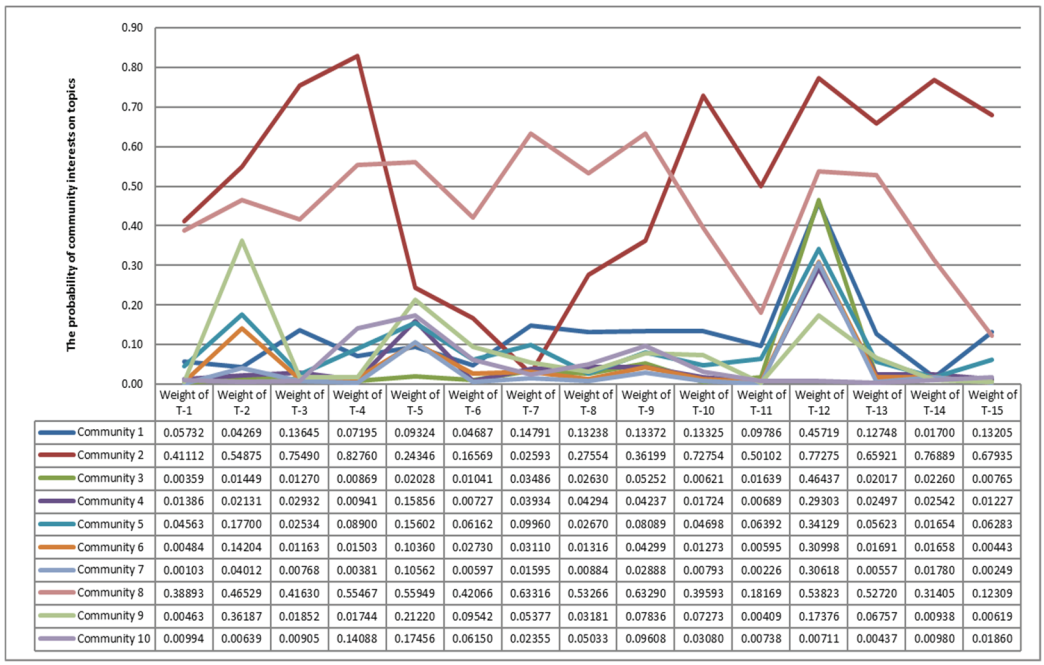


Fig. 5. Visualizing results of exploring interested "Project" (Dự án) topic with community (cộng đồng).

The table in Fig. 6 offers insights into the weight vectors and communities for each topic. One striking point is that community 2 has the highest probability for multiple topics, from T-1 to T-15, while the numbers of community 8 are moderately smaller. There are no remarkable differences in the probabilities of community 3, 7, and 10 for the 15 topics above, with figures ranging from 0.00103 to 0.17376. However, community 10 at the weight of T-13 has the smallest probability of 0.00437.



	Weight of T-1	Weight of T-2	Weight of T-3	Weight of T-4	Weight of T-5	Weight of T-6	Weight of T-7	Weight of T-8	Weight of T-9	Weight of T-10	Weight of T-11	Weight of T-12	Weight of T-13	Weight of T-14	Weight of T-15
Community 1	0.05732	0.04269	0.13645	0.07195	0.09324	0.04687	0.14791	0.13238	0.13372	0.13325	0.09786	0.45719	0.12748	0.01700	0.13205
Community 2	0.41112	0.54875	0.75490	0.82760	0.24346	0.16569	0.02593	0.27554	0.36199	0.72754	0.50102	0.77275	0.65921	0.76889	0.67935
Community 3	0.00359	0.01449	0.01270	0.00869	0.02028	0.01041	0.03486	0.02630	0.05252	0.00621	0.01639	0.46437	0.02017	0.02260	0.00765
Community 4	0.01386	0.02131	0.02932	0.00941	0.15856	0.00727	0.03934	0.04294	0.04237	0.01724	0.00689	0.29303	0.02497	0.02542	0.01227
Community 5	0.04563	0.17700	0.02534	0.08900	0.15602	0.06162	0.09960	0.02670	0.08089	0.04698	0.06392	0.34129	0.05623	0.01654	0.06283
Community 6	0.00484	0.14204	0.01163	0.01503	0.10360	0.02730	0.03110	0.01316	0.04299	0.01273	0.00595	0.30998	0.01691	0.01658	0.00443
Community 7	0.00103	0.04012	0.00768	0.00381	0.10562	0.00597	0.01595	0.00884	0.02888	0.00793	0.00226	0.30618	0.00557	0.01780	0.00249
Community 8	0.38893	0.46529	0.41630	0.55467	0.55949	0.42066	0.63316	0.53266	0.63290	0.39593	0.18169	0.53823	0.52720	0.31405	0.12309
Community 9	0.00463	0.36187	0.01852	0.01744	0.21220	0.09542	0.05377	0.03181	0.07836	0.07273	0.00409	0.17376	0.06757	0.00938	0.00619
Community 10	0.00994	0.00639	0.00905	0.14088	0.17456	0.06150	0.02355	0.05033	0.09608	0.03080	0.00738	0.00711	0.00437	0.00980	0.01860

Fig. 6. Weight vector (w_i) has a component representing the level of interested topic of each community in January 2019 (T stands for Topic).

5. Evaluation and Discussion of Results

5.1. Comparison of the Kohonen Network to K-Medoids Clustering Method

Apart from applying the Precision, Recall, and F1 scores to evaluate the test results, this article employs the root mean square standard deviation (RMSSTD) and R-squared (RS) values [26,27] to compare the results of the clustering method proposed in the paper with those of the algorithm for K-Medoids. The RMSSTD value is used to measure the quality of the collection algorithm by formula (2). For RMSSTD, a lower value indicates a better clustering.

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

where k is the number of clusters, p is the number of independent variables in the dataset, $\overline{x}_{ij}$ is the average of data of variable j, and cluster i, n_{ij} is data in variables p and k clusters.

The average of RMSSTD is calculated based on 1,000 transactions for each dataset. Formula (3) calculates the average value of RMSSTD:

$$\text{RMSSTD average} = \frac{\textit{The total value} \text{ of RMSSTD from 1,000 transactions for which the dataset was performed}}{1{,}000} \tag{3}$$

The RS value is used to consider significant differences in data objects between different clusters and in a highly similar cluster. If the RS value is 0, then there is no difference between clusters. In contrast, if the RS value is 1, then the clustering result is optimal. The RS value is calculated using formulas (4), (5), and (6):

$$RS = \frac{SS_t - SS_w}{SS_t} \tag{4}$$

$$SS_t = \sum_{j=1}^{p} \sum_{a=1}^{n_j} (x_a - \overline{x}_j)^2 \tag{5}$$

$$SS_w = \sum_{j=1}^{i=1..k} \sum_{a=1}^{n_{ij}} (x_a - \overline{x}_{ij})^2 \tag{6}$$

where SS_t is the sum of squares of distances between all variables, SS_w is the sum of squares of distances between all data objects in the same cluster, where k is the number of clusters, p is the number of independent variables in the dataset, $\overline{x}_{ij}$ is the data average of variable j and cluster i, n_{ij} is the amount of dataset in variable p and cluster k.

The average value of the RS is calculated based on 1,000 iterations of each dataset being performed. This value is calculated by formula (7).

$$\text{RS Average} = \frac{\textit{The total value} \text{ of RS from 1,000 iterations of each dataset}}{1{,}000} \tag{7}$$

5.2 Evaluation of Experimental Results and Discussion

Evaluation by RMSSTD and RS values

The dataset, which consists of vector sets from the results of the TART model (Table 1), and evaluation methods are used to test clustering methods to find the average values of RMSSTD and RS. The test was repeated 1,000 times to obtain stable, reliable results. The number of *k* clusters has also been changed to have more criteria for comparing different methods.

Table 2 shows average RMSSTD values. The Kohonen neural network method has lower RMSSTD values than the K-Medoids method. This means that the Kohonen neural network has a better performance than the K-Medoids algorithm.

In this experiment, two clustering algorithms are compared using RMSSTD and RS values of the actual dataset from the TART model results. The calculation shows that the Kohonen neural network algorithm yields the lowest RMSSTD values and the highest RS values. This indicates the Kohonen algorithm is better than the others. This can be explained by the fact that the datasets used in this research do not include noise or outlier data.

Table 2. Average results of RMSSTD and RS for two clustering methods

Cluster (*k*)	RMSSTD		RS	
	Kohonen	K-Medoids	Kohonen	K-Medoids
2	**0.56032**	0.67832	**0.67632**	0.61231
3	**0.65235**	0.76234	**0.68932**	0.62311
4	**0.57642**	0.65231	**0.74350**	0.65634
5	**0.54324**	0.58932	**0.72341**	0.66549
6	**0.46352**	0.49812	**0.79831**	0.75410
7	**0.49482**	0.57321	**0.87322**	0.81209
8	**0.41521**	0.46421	**0.84321**	0.78619

Evaluation by Precision, Recall, and F1-score

The precision between the two clusters, denoted *P*, reflects the query's accuracy and is calculated using formula (8). The precision indicates the ratio between the number of correctly clustered messages. If *P* = 1, then the messages in cluster k_i are in the messages of cluster m_i. Given precision *P*, *a* is the common part of two comparison clusters *b* and *c* [28].

$$P = \frac{a}{a + c} \tag{8}$$

Recall [28] between two clusters m_i and k_i is denoted R and calculated by formula (9). If $R = 1$, the messages in cluster m_i belong to messages in cluster k_i:

$$R = \frac{a}{a + b} \tag{9}$$

Combining precision with recall yields the F1 score [27]:

$$F = \frac{2PR}{P + R} \tag{10}$$

According to Brew and im Walde [29], the evaluation method is as follows. First, corresponding to one cluster in the clustering result, the system will calculate the value of the F1-score with all clusters collected manually. The next step is picking out the highest F1-score and removing this cluster. The process continues the calculation for the rest of the clusters. The higher the total F1-score is, the more accurate it is for the cluster method. Below are the results of the F1-score corresponding to Table 3 during March 2019 and April 2019, with $m = 5$ clusters and $k = 6$ clusters.

The total MAX values of the F1-score in Table 3 are 3.77 compared to 5 and 4.08 compared to 5 during March 2019 and April 2019, respectively. This max value is high, proving the effectiveness of the community discovery (clustering) method proposed in the article by combining the Kohonen neural network method and the TART topic model to achieve high efficiency.

Table 3. Results of comparing the F1 scores of manual and machine clustering (Kohonen) for March 2019 and April 2019

Kohonen (k)	**Manual (m)**									
	March 2019					**April 2019**				
	m_0	m_1	m_2	m_3	m_4	m_0	m_1	m_2	m_3	m_4
k_0	0.52	0.46	0.32	0.72	0.78	0.43	0.67	0.76	0.47	**0.78**
k_1	**0.85**	0.71	0.19	0.54	0.32	**0.84**	0.43	0.47	0.39	0.00
k_2	0.00	0.65	0.58	**0.81**	0.00	0.45	**0.79**	0.34	**0.85**	0.35
k_3	0.79	0.00	**0.72**	0.23	0.54	0.72	0.00	0.00	0.52	0.62
k_4	0.56	0.42	0.16	0.00	**0.82**	0.29	0.78	**0.82**	0.63	0.48
k_5	0.52	**0.76**	0.00	0.29	0.21	0.00	0.45	0.21	0.00	0.31
MAX	**0.85**	**0.76**	**0.72**	**0.62**	**0.82**	**0.84**	**0.79**	**0.82**	**0.85**	**0.78**

6. Conclusion and Future Work

This study makes three important and practical scientific contributions to user experience and community discovery.

First, the topic model was applied to social network analysis to discover topics from messages on social networks. The paper proposes a method that combines the topic model with labeling based on topic taxonomy. This method serves as the foundation for further research on the discovery, content analysis, and labeling to offer fresh insights into users' experiences through social networks.

Second, this article shows how the TART model can be applied to assess the role of the individual's interest in a topic based on a temporal factor. This model plays an essential role in finding the relationship among individuals on social media within the topic model. The model's output is a set of vectors that consist of individuals' traits on social networks.

Third, a method was constructed and developed to discover the community's interests using the TART topic model. This method helps to identify the groups of users who have the same topic of interest but whose degree of interest varies across topics for each period. In addition, the Kohonen neural network is trained to discover the community of users interested in each topic. This proposal is called the community discovery method based on the TART topic model and clustering method. In particular, the method of community discovery distributes topics by community, specific topics of interest, and their probabilities. The results of community discovery are visualized in the Kohonen output layer.

In future work, we will concentrate on analyzing the impact of the community's topic spread on social networks. This analysis will aim to determine the path and the source of information. Further, we can build time systems (containing the overlap property) to analyze online social networks for different periods with the topic model and the big data solution. The model of discovering behaviors and customers' experience in the tourism sector is pivotal, based on the big data platform and the topic model employed in this setting.

References

[1] C. C. Aggarwal, *Social Network Data Analytics*. Boston, MA: Springer, 2011.

[2] L. Berkani, S. Belkacem, M. Ouafi, and A. Guessoum, "Recommendation of users in social networks: A semantic and social based classification approach," *Expert Systems*, article no. e12634, 2020. https://doi.org/10.1111/exsy.12634

[3] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in *Proceedings of the 2012 SIAM International Conference On Data Mining*, Anaheim, CA, 2012, pp. 624-635.

[4] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, and X. Li, "The author-topic-community model for author interest profiling and community discovery," *Knowledge and Information Systems*, vol. 44, no. 2, pp. 359-383, 2015.

[5] D. Zhou, I. Councill, H. Zha, and C. L. Giles, "Discovering temporal communities from social network documents," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, Omaha, NE, 2007, pp. 745-750.

[6] N. Pathak, C. DeLong, K.Erickson, and A. Banerjee, "Social topic models for community extraction," *Department of Computer Science and Engineering, University of Minnesota, Minneapolis*, MN, 2008.

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

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

[9] A. Beykikhoshk, O. Arandjelovic, D. Phung, and S. Venkatesh, "Discovering topic structures of a temporally evolving document corpus," *Knowledge and Information Systems*, vol. 55, no. 3, pp. 599-632, 2018.

[10] L. C. Freeman, "Visualizing social networks," *Journal of Social Structure*, 2000 [Online]. Available: https://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

[11] H. H. Kim and H. Y. Rhee, "An ontology-based labeling of influential topics using topic network analysis," *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1096-1107, 2019.

[12] Z. Yin, L. Cao, Q. Gu, and J. Han, "Latent community topic analysis: Integration of community discovery with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, pp. 1-21, 2012.

[13] T. Ho and P. Do, "Analyzing the changes in online community based on topic model and self-organizing map," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 6, no. 7, pp. 100-108, 2015.

[14] D. M. Sharma and M. M. Baig, "Sentiment analysis on social networking: a literature review," 2015 [Online]. Available from: https://www.researchgate.net/profile/Durgesh_Sharma8/publication/325120893_Using_Data_Mining_For_Prediction_A_Conceptual_Analysis/links/5ef35b3d92851c35353ba7c4/Using-Data-Mining-For-Prediction-A-Conceptual-Analysis.pdf.

[15] H. Fani, F. Zarrinkalam, X. Zhao, Y. Feng, E. Bagheri, and W. Du, "Temporal identification of latent communities on Twitter," 2015 [Online]. Available: https://arxiv.org/abs/1509.04227.

[16] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 306-315.

[17] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks: a Bayesian approach," *Machine Learning*, vol. 82, no. 2, pp. 157-189, 2011.

[18] T. Griffiths, "Gibbs sampling in the generative model of latent Dirichlet allocation," 2002 [Online]. Available: https://people.cs.umass.edu/~wallach/courses/s11/cmpsci791ss/readings/griffiths02gibbs.pdf.

[19] J. Singh and A. K. Singh, "NSLPCD: topic based tweets clustering using node significance based label propagation community detection algorithm," *Annals of Mathematics and Artificial Intelligence*, 2020. https://doi.org/10.1007/s10472-020-09709-z

[20] T. Ho and P. Do, "Social network analysis based on topic model with temporal factor," *International Journal of Knowledge and Systems Science (IJKSS)*, vol. 9, no. 1, pp. 82-97, 2018.

[21] H. A. Abdelbary, A. M. ElKorany, and R. Bahgat, "Utilizing deep learning for content-based community detection," in *Proceedings of 2014 Science and Information Conference*, London, UK, 2014, pp. 777-784.

[22] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, 1982.

[23] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999. pp. 443-465.

[24] Kohonen T, "Self-Organization and Associative Memory", *Springer*, Berlin, 1984.

[25] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 16th International Conference on Machine Learning (ICML)*, Bled, Slovenia, 1999, pp. 200-209.

[26] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: part I," *ACM SIGMOD Record*, vol. 31, no. 2, pp. 40-45, 2002.

[27] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering validity checking methods: Part II," *ACM SIGMOD Record*, vol. 31, no. 3, pp. 19-27, 2002.

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[29] C. Brew and S. S. im Walde, "Spectral clustering for German verbs," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, 2002, pp. 117-124.

Thanh Ho (Ho Trung Thanh) **https://orcid.org/0000-0002-9033-3735**

He received M.S degree in computer science from University of Information Technology, VNU-HCM, Vietnam in 2009 and Ph.D. degree in computer science from University of Information Technology, VNU-HCM, Vietnam. Dr. Ho is currently lecturer in Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam. His research interests are data mining, data analytics, business intelligence, social network analysis and big data.

Tran Duy Thanh **https://orcid.org/0000-0003-0680-9452**

He received his M.S. degree in computer science from University of Information Technology, VNU-HCM, Vietnam. Mr. Tran is currently lecturer in Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam. His research interests are social network analysis, big data, AI and robotics.