Analyzing the Changes in Online Community based on Topic Model and Self-Organizing Map

Thanh Ho
Faculty of Information System, University of Economics and Law, VNU-HCM
Ho Chi Minh City, Vietnam

Phuc Do
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam

***Abstract*—In this paper, we propose a new model for two purposes: (1) discovering communities of users on social networks via topics with the temporal factor and (2) analyzing the changes in interested topics and users in communities in each period of time. This model, we use Kohonen network (Self-Organizing Map) combining with the topic model. After discovering communities, results are shown on output layers of Kohonen. Based on the output layer of Kohonen, we focus on analyzing the changes in interested topics and users in online communities. Experimenting the proposed model with 194 online users and 20 topics. These topics are detected from a set of Vietnamese texts on social networks in the higher education field.**

***Keywords*—SOM; topic model; interested topics; online users; online community; social networks**

I. INTRODUCTION

In the scope of this paper, we would like to mention users' community on social networks. Online community on the social network is a group of individuals who interact through specific media are able to overcome geographical boundaries and politics to pursue common interests or goals. In [5][10][12][16][20], community is a group of users who live and work in the same environment. One of the most popular virtual community types is social networking community.

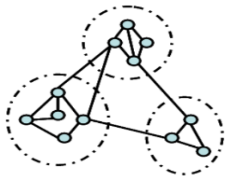

Fig. 1. Community on social networks [5]

Figure 1 shows the structure of a social network through interacting of users in communities [5]. There are 3 communities and the links are among communities by discussed messages. It can be defined the community is a group of users on the social network who have interaction with each other and often pay attention to the discussed topics in the group rather than other groups [15][16][18]. In this paper, the set of online communities is denoted with C and a community is denoted with c, $c \in C$.

The conditional probability of a user community represents levels of participation and interested topics of users in communities [18]. In particular, $p(c|u)$ is the probability of community c that contains the user u [5]. Thus, each user u is contained in only the community in each period of time. In our research, we don't consider overlap community. We consider discrete communities on topics. For example, we have a list of communities: $\{C_1, C_2, C_3, C_4,...,C_k\}$ (1)

Users' interests in topics often changes. This makes online communities of users also change. The influence leads to changes in online communities with two major reasons: (1) from the formation or change of groups of acquainted friends who make friends online or via the introduction of friends; (2) through hobbies of online users who make friends with each other or users who are interested in topics in message contents that users discuss. Thus, the relationship of online communities is regarded as a network with the combination of users. This relationship is shown on social networks [4][5][7][29]. Because of properties of each user on the social network, message contents exist in form of texts, images, etc. In a period of time, the same online community could be interested in exchanging many topics, and a topic can be discussed by many communities. Our research tasks are how to discover online communities of users on topics of messages discussed by users in communities and how each community is interested in a specific topic.

Another challenge given is that online communities often changes components on social networks over time, such as changes of users in communities, interested topics, etc. Therefore, the components changing in communities are often relevant to one or many topics that communities notice on social networks, the number of users in communities, levels of interests in each topic over time, and more particularly, changes in online community that have a lot of influences on behavior, attention and exchanges of users in communities. This leads to attracting many researchers paying attention to analyzing and facing the spread information to find out the origin of the sender's information [15][27] or discover the influence of users or important topics to serve development strategies, such as managing users in companies, educational organizations or a country with the purpose of understanding users and performing effective marketing strategies, orientating careers and improving training environment, etc.

In order to discover the community of users on topics in each period of time, in this paper, we approach the topic model to exploit possibilities of content analysis to find each topic in each message content along with a specific set of words according to topics [4][8][9][25][26] and continue to exploit efficiency of our TART model to discover communities on interested topics of users with the temporal factor we propose and introduce in the study [22].

Besides the effective exploitation of TART model [22], we propose models that explore the community of users on the social network by using the training method of Kohonen network [6][21][23] combined with TART model. Subsequently, we focus on analyzing the change of topics and users of the community in each period of time.

The next sections of the paper: section 2 presents the related researches; section 3 presents the proposed model that discovers the community of users on the social network and analyzes the change of interested topics of users of communities in each period of time; section 4 presents experimental results and evaluation; section 5 presents conclusion, development directions and references.

II. Related Works

A. Group-Topic Model (GT)

In [7], authors aim to use as much of the commonly shared information that is available for the purposes of entity resolution. This information is organized via the latent concept of a group of authors (which characterizes which authors might be co-authors) along with topic information associated with each group (which helps disambiguate authors which could be authors of a number of groups). This leads to a model which authors call the grouped author-topic model.

To describe the model we need to introduce two concepts, that of group and that of topic. The idea of topic is common to other papers on topic model, where a topic is a mixture component defining a distribution of words. An individual abstract will only contain a small number of topics out of the total possible number. This is a result of the model taking a Bayesian non-parametric approach to the problem and allowing broad uninformative priors to be set on the number of entities.

B. Community-User-Topic model (CUT)

In [29], the authors propose two generative Bayesian models for semantic community discovery in social networks, combining probabilistic modeling with community detection in social networks. To simulate the generative models, an Gibbs sampling algorithm is proposed to address the efficiency and performance problems of traditional methods. In which, [29] approach successfully detects the communities of individuals and in addition provides semantic topic descriptions of these communities with two models: CUT_1 and CUT_2. CUT_2 differs from CUT_1 in strengthening the relation between community and topic. In CUT_2, semantics play a more important role in the discovery of communities. Similar to CUT_1, the side-effect of advancing topic z in the generative process might lead to loose ties between com-munity and users

C. Community-Author-Recipient-Topic model (CART)

In [5], the authors introduce CART model (Community - Author - Recipient - Topic), the model is tested on the Enron email data system[1]. The model shows that the discussion and exchange between users within a community are related to the other users in community. This model is binding on all relevant users and the topics discussed in the emails belonging to a community, while the same users and the various topics can link to other communities. Compared with the above models including CUT, CART model is closer to further emphasize the ways that the topics and their relationships affect the structure of the online community in exploring community on topics.

[1] https://www.cs.cmu.edu/~./enron/

CART model [5] is one of the first attempts to discover the community by combining research-based content message that users of community to exchange on social network. The model consists of 4 main components in CART are C, A, R and T. In particular, C is a community of users, R is the recipients, A is authors, T is topics [5].

The CART model has the following:

1) To generate email e_d , a community c_d is chosen uniformly at random.

2) Based the community c_d, the author a_d and set of recipients ρ_d are chosen.

3) To generate every word $w_{d,i}$ in that email, a recipient $r_{d,i}$ is chosen uniformly at random from the set of recipients ρ_d.

4) Based on the community c_d, author $a_{d,}$ and recipient $r_{d,i}$, a topic $z_{d,i}$ is chosen.

5) The word $w_{d.i}$ itself is chosen from the topic $z_{d,i}$.

Gibb sampling for CART model as:

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

where, ρ_d set of recipients R, r_d is the sequence of latent recipients (selected from ρ_d), a_d is author and z_d is the sequence of latent topic corresponding to word sequence $w_{d.i}$ in document d, and N_d is the total number of words in the email.

III. Motivation Research

The above-mentioned studies [3][5][7][29] and other studies such as [3][10][23][24][30] studied the models of discovering communities based on content analysis. However, these studies have not attached special importance to the temporal factor and analyzed the changes in users' interests in topics in community in each period of time. Because the changes in users' interests in topics can affect changes in interested topics of communities and may also change the components of the online community, such as the geographical area forming community, the number of users, time and topics in community. We focus on analyzing the distribution of interested topics in the online community and analyzing the changes in interested topics and users in communities.

IV. Discovering Community Model

A. Kohonen network

Kohonen network was invented by a man named Teuvo Kohonen, a professor of the Academy of Finland. The Self-Organizing Map (SOM), commonly also known as Kohonen network is a computational method for the visualization and analysis of high-dimensional data, especially experimentally acquired information [2][17][19][28].

Determine the suitability through the survey of relevant researches and use of methods and algorithms for clustering to explore communities of users on topics, we choose the method Kohonen network. Kohonen network can cluster data without prior designated clusters (cluster correlation data in this study are interested topics of online community, corpus message

enormous, multi-dimensional and online community very large should the predetermined number of clusters is extremely difficult) [17][19][23]. In addition, the output layer of Kohonen network is capable of performing visual text blocks, topics through the Kohonen layer in 2D [13][17][19].

The goal of the Kohonen network is mapped to N-dimensional input vector into a map with 1 or 2 dimension [2][3][19][20][28]. The vector space together in input will close on output layer of Kohonen network. A Kohonen network consists of a grid of the output node and the input node N. Vector input is transferred to each output node (see figure 2). Each link between input and output of Kohonen network corresponds to a weight. Total input of each neuron in the Kohonen layer by total weight of the input neurons that.

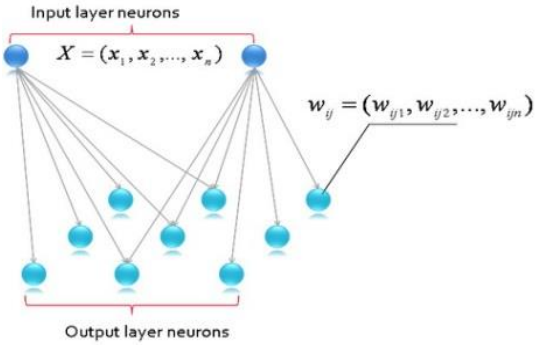

Fig. 2. The Kohonen neuron network structure for clustering vectors [3]

In initializing input and output layers, according to the figure 2, the input layer is a unique vector X .Each dimensional value of X such as x_1, x_2 or x_n is represented as a certain input layer neurons in the figure 2. On the other hand, output layers of Kohonen network is a three-dimensional matrix of neurons. The self-organizing map is described as a square matrix since each output layer neurons is a group of one-dimensional matrix or a vector of weights with the number of its element is the number of input layer neurons or the number of dimensions of input vector - n. Therefore, the data we need for initializing input and output layer neurons will be:

- Let n be the number of dimensions of the input vector or the number of interested topics.
- And m be the number of elements for the output layer or the self-organizing map.

We use input vectors as in table 1 and table 2, in this case, n is equal to 3. Because these vectors have 3 dimensions or interested topics and m is depend on how many output neurons. As a result, output neurons are a SOM with m element and each element has 3 weights or we have m vectors in the output neuron layers. The reason for this outcome is in the learning process from each vector of learning set, we need to find the winning output neuron then we updates the value for relevant neurons which depends on the winning neuron and the current input vector (see figure 3).

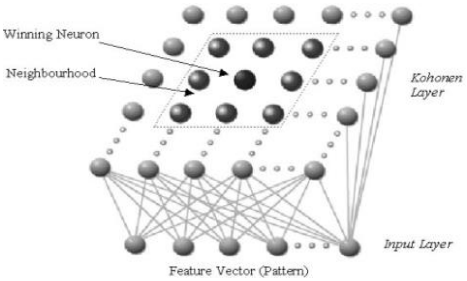

Fig. 3. Finding winning neuron and its neighborhood[2]

The winning neuron is determined by finding the shortest distance neurons in the set of results. After winning neuron identified, the next step determines the vicinity of the winner neuron. The algorithm will update the weights of the weight vector of the winning neuron and all the neurons located in the neighborhood of the winner neuron. To determine the vicinity of winning neuron (called winning region), neighborhood function is applied. The function is described as follows:

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

where, r is the distance between w_x (a winning neuron vector) and w_i (a current neuron vector)

$$r = \sqrt{(i - i_0)^2 + (j - j_0)^2} \quad (4)$$

where i_0, j_0 are ordinate of winning neuron vector and $\sigma(t)$ is the function for identifying the space of the neighborhood. In the beginning of the function, it involves almost the whole space of the grid, but with time, the value of σ decreases [1].

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

with:

- $\alpha(t)$: the learning rate at the iteration t
- α_0: the initializing value of learning rate, $\sigma_0 = \sqrt{m}$
- t: the current number of iterations
- τ_1: constant

The neighborhood function is represented as:

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

Use Mexican hat function to identify the neighborhood of winning neuron for the input vector. To be more understandable and comprehensible, the formula for updating weight showed as follows:

$$w'_{(i,j)_k} = w_{(i,j)_k} + \alpha(t)h(r,t)\left(v_{x_k} - w_{(i,j)_k}\right) \quad \forall k \in \mathbb{N}, 0 \le k \le n \quad (7)$$

where,

- k: the dimension of neuron weights
- n: the number of interested topics
- $w'_{(i,j)_k}$: the new value (post-update) of k[th] weight of the neuron at row i, column j

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

- $w_{(i,j)_k}$: the current (pre-update) value of k[th] weight of the neuron at row i, column j
- $\alpha(t)$: the learning rate at the current number of iterations
- $h(r,t)$: the result of topological neighborhood function with t is the current number of iterations, r is the distance between the current neuron and the winning neuron
- v_{x_k}: the value of k[th] weight of the current learning vector v_x

Function α(t) is the learning rate, this value will decrease as the number of iterations t. If a neuron is a winning neuron or neighborhood of the winner neuron, then the weight of vector is updated, reverse that neuron will not be updated. At each iteration, SOM have chosen the same weight vector to update its vector and weight vector to make them closer to the input vector.

B. Temporal – Author – Recipient – Topic model (TART)

We proposed a Temporal-Author-Recipient-Topic model [23] in the field of social network analysis and information extraction based on the topic model. The key ideas of the model focus on extracting words, discovering and labeling topics, and analyzing topics with authors, recipients and temporal factor.

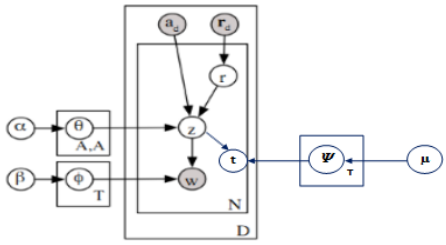

Fig. 4. TART model [23]

During parameter estimation for TART model, the system will keep track 4 matrices to analyze users' interests topics, including: T (topic) x W (word), A (author) x T (topic), R (recipient) x T (topic) and T (topic) x T (temporal). Based on these matrices, topics and temporal distribution Φ_{zw}, topic and temporal distribution Ψ_{zt}, author and topic distribution Θ_{az}, recipient and topic distribution Θ_{rz}, the matrices are given by (8), (9), (10) and (11):

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

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

C. General model

We propose the model for discovering online community and analyzing the changes in topics interests and users in communities on social networks in each period of time approaching the topic model with temporal factor. In this model, through results of the analysis and evaluation of the relevant models in discovering communities, we choose Kohonen network. Kohonen network combines with TART model [23]. The output of TART model is the set of interested topic vectors of users in each period of time. The general model consists of 3 main modules (figure 5):

1) Normalization the set of vectors from the output of TART model in order to suit the input vectors of Kohonen network.

2) Discovering community by using Kohonen network (SOM) to cluster users based on interested topic vectors. In this discovery, each cluster is a community of users on topics, corresponding to a neuron on the output layer of SOM.

3) Analyzing the changes of users and interested topics in communities on social networks based on the output layer of SOM and the relationship among output layers.

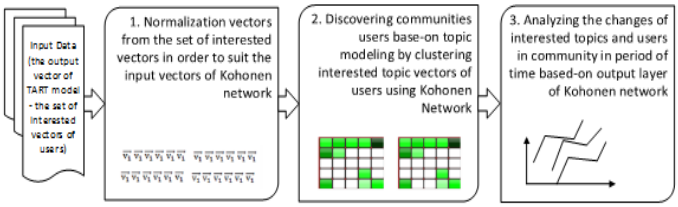

Fig. 5. General model of discovering community

The algorithm 1 describes the way to discover community users based-on topic model combined Kohonen network by clustering interested topics vectors of users and analyze the changes in communities of users.

Algorithm 1. Discovering communities and analyzing the changes in communities of users.

Input: the set of interested topic vectors of users (called the set of input vectors) from TART model [22]. The components of vectors include the topics probability and temporal factor which users are interested in.

Output: the set of communities of users on specific topics in each period of time and the changes in interested topics and users in online communities.

Process: Using method of Kohonen network. In this method, we introduce the main process steps, include:

1. Putting the set of input vectors.
2. For each i ϵ [1,...,n] //n is row and column on output layer of Kohonen.
   For each j ϵ [1,...,n]
   Finding neurons which have weight vectors w_{ij} nearest with input vector v. Called (i_0, j_0) is of winning neuron. Hence, euclician distance between $d(v,w_{i0j0}) = \min (d(v,w_{ij})$ with i,j ϵ [1,...,n] and w_{i0j0} are weight of winning neuron.
3. Finding winning neuron and its neighborhood (figure 3)
4. Discovering online community based on the winning neuron and its neighborhood.
5. Analyzing the changes in interested topics and users in online communities based on online community on the output layer of SOM.

V. IMPLEMENTATION AND DISCUSSION

A. Experimental Data

Experimenting the proposed model (figure 5) for discovery communities with 194 interested topic vectors of 194 users who discuss 9 topics (random survey 9 topics are "facilities and services", "learning and examination", "international cooperation", "quality control", "scientific research", "living and life", "sport", "employment recruitment", "admission", "finance and fees", "friendship and love", "social activities" and "training" from 20 topics in the system of topics built in [11]). We analyze the above topics belonging to the period from December, 2008 to January, 2010 on 48.264 messages from social networks. In each period of time, we have interested topic vectors of different users. For example, the user u_1 during the period from t_1 to t_2, has an interested topic vector of user $v(u_1, t_1, t_{2)}, u_1 \in U$, during the period from t_2 to t_3, we have the vector $v(u_1, t_2, t_{3)}$. In general, each user has an interested topic vector at the time t is $v_i(t) = < v_{i_1}^t, v_{i_2}^t, v_{i_3}^t, \dots, v_{i_n}^t >$ or $X = (x_1, x_2, \dots, x_n)$. Thus, we have interested topic vectors of users as follows:

TABLE I. THE SET OF INTERESTED TOPIC VECTORS

User	Temporal t_i	Temporal t_j	$v(u, t_i, t_{j)}$
u_1	Dec 01, 2008	Dec 31, 2008	$v(u_1, t_1, t_{2)}$
u_2	Feb 01, 2009	Feb 28, 2009	$v(u_2, t_2, t_{3)}$
u_3	Apr 01, 2009	Apr 30, 2009	$v(u_3, t_3, t_{4)}$
u_1	Feb 01, 2009	Feb 28, 2009	$v(u_1, t_2, t_{3)}$

TABLE II. THE SET OF INTERESTED TOPIC VECTORS OF USERS IN OTHER FORM

Users	Topic "international cooperation"	Topic "admission"	Topic "learning and examination"	Temporal $t_i - t_j$
	Interested Probability			
u_1	0.85246	0.0	0.772527	Dec 01, 2008 - Dec 31, 2008
u_2	0.85000	0.86956	0.676793	Feb 01, 2009 - Feb 28, 2009
u_3	0.62417	0.34132	0.893421	Apr 01, 2009 – Apr 30, 2009
u_1	0.52345	0.52341	0.834212	Feb 01, 2009 - Feb 28, 2009

Table 1 and table 2 are the forms of interested topics of users on social networks. This is the set of input vectors for Kohonen network. The input vectors include 3 users interested in 3 topics in 3 periods of time t_1-t_2, t_2-t_3 and t_3-t_4. The goal of training process is to cluster the set of interested topic vectors. Thus, with $V(t_i, t_{j)}$ we have the output layer of Kohonen $K(t_i, t_j)$ which is a 2-dimensional array (see figure 9).

B. Discovering online community

This section presents the results of test to discover communities of users on the social network in each period of time. This section focuses on modules (1) and (2) of the model in figure 6.

Figure 6 shows the results of the training process to discover communities of users on the output layer, experimenting with 194 topic vectors with 194 users in discussing on 9 topics.

Each neuron (cell) on the output layer (see figure 6) corresponds to a community of users to exchange topics in each period of time. Each neuron has a dark or light color corresponding to the number of users more or less in communities. The darker the color on each neuron is, the more the number of users in the community is. If the neuron is white, users in communities do not exist.

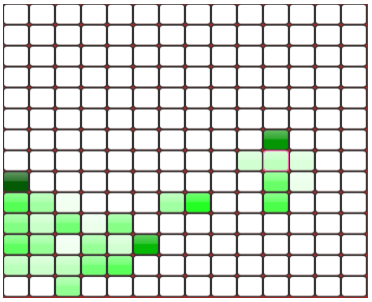

Fig. 6. Results of discovery communities is shown on the output layer of SOM

C. Analyzing the changes in interested topics and users in online communities

This section focuses on testing the proposed model of the module (3) in figure 5. Based on the output layer of SOM in each period of time in figure 6, we can examine the relationship between the clusters (neurons) in the output layer based on the components such as users, interested topics, probability and number of clusters in each period of time.

Based on the output layer of SOM in each period of time in figure 6, we can examine the relationship between the clusters (neurons) in the output layer based on the components such as users, interested topics, probability and number of clusters in each period of time.

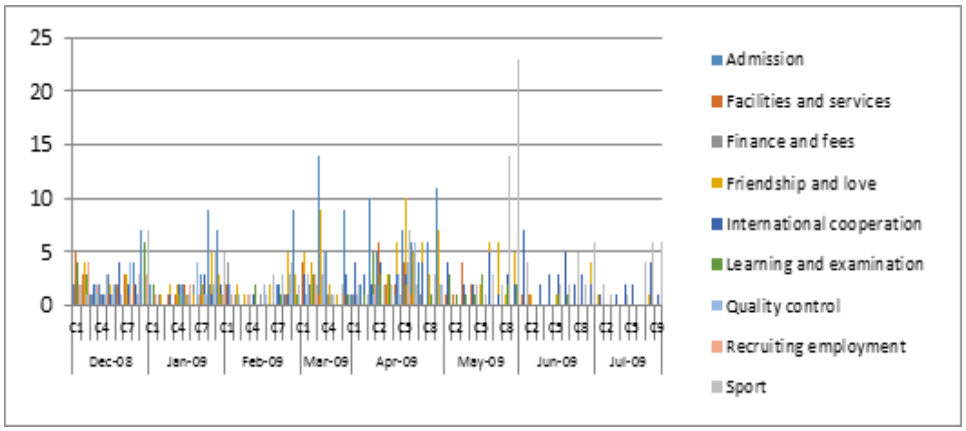

Fig. 7. Analyzing the changes in interested topics in community of users in each period of time

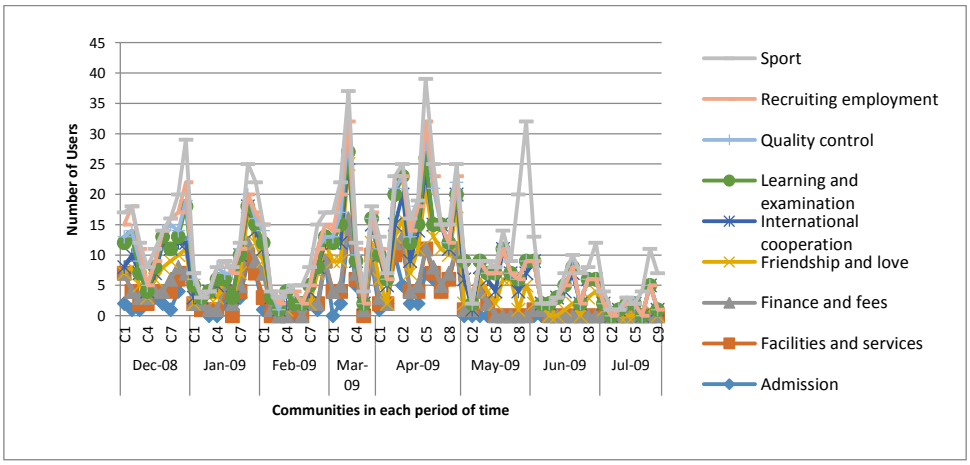

Fig. 8. Analyzing the changes in interested topics in community of users in each period of time

Figure 7 and figure 8 show the analyzed results of changes in interested topics and users in the communities from Dec-2008 to Jul-2009. Surveying 9 topics, we find that interested topics have frequent levels during months and highly increase in Apr-2009 and May-2009, and occupy most users in communities with 9 topics. Besides, we find that interested topics have frequent levels during months and highly decrease in Jun-2009 and Jul-2009.

Figure 9a, 9b and 9c show the output layer of Kohonen in 3 periods of time (Mar-2009, Apr-2009 and May-2009). We have the output layer with a set of neurons (each neuron in dark color is the one corresponding community of users on specific topics).

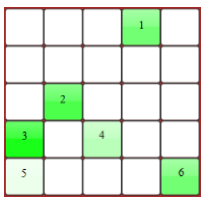

Figure 9a. Results of discovery 6 communities show on the output layer in Mar-2009.

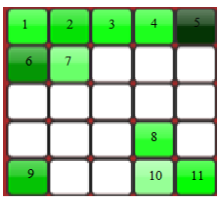

Figure 9b. Results of discovery 11 communities show on the output layer in Apr-2009.

Figure 9c. Results of discovery 9 communities show on the output layer in May-2009

Fig. 9. Results of discovery communities show on the output layer of Kohonen

Based on the output layer in figure 6 and figure 9 (9a, 9b, 9c), we continue to analyze changes in interested topics and users in communities in each period of time). Each period of time, there are different from the number of communities among output layers of Kohonen. In figure 9a, there are 6 communities. However, figure 9b has 11 communities in Apr-2009 and figure 9c has 9 communities in May-2009. Figure 10 shows the changing of topics interested in communities on output layers of Kohonen (figure 9).

According to figure 11, there are 3 communities C_5, C_6 and C_7 on 9 topics. At that time, community C_5 is interested in the 8 topics in May-09. In each period of time, the participation level of users in communities on topics may also change in other communities. Observing figure 11, we see the elasticity of the number of users in each community in each period of time. In this observation, the community C_2 in the topic "learning and examination" in Dec-2008 with the number of users is 16, but in Jan-2009, the number of users in community C_3 is 4, in Jun-2009 is 2, but in Jul-2009, community in the topic "learning and examination" doesn't exist anymore. Analyzing the data, we find that during Jul-2009, most users are only interested in the topic "international cooperation" in community C_9.

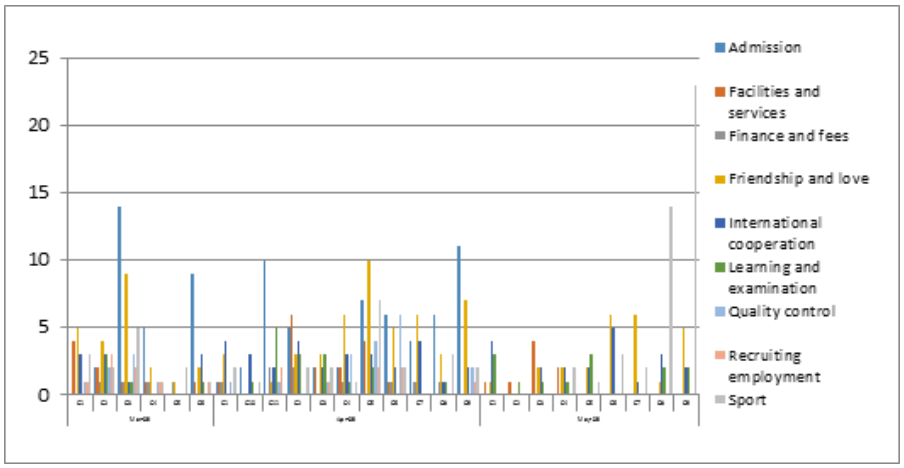

Fig. 10. Communities on 9 topics in 3 periods of time (Mar-2009, Apr-2009 and May-2009) on the output layer of SOM

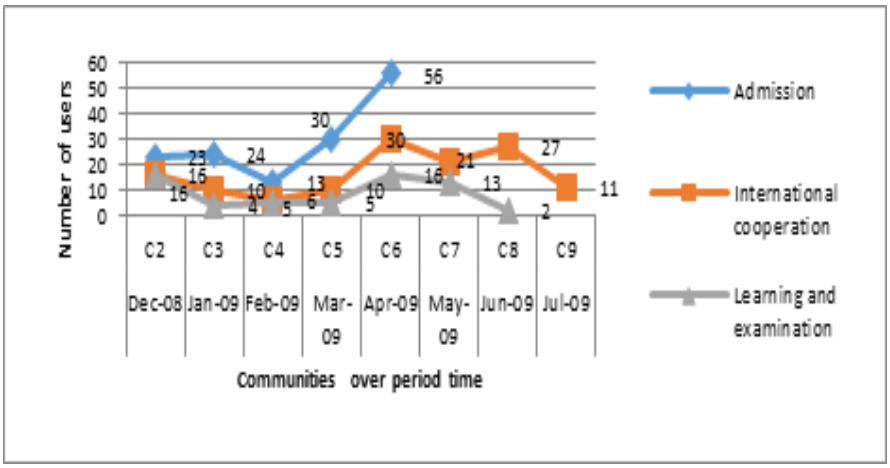

Fig. 11. The change users of communities on topic in period from Dec-2008 to Jul-200

However, in Feb-2009, the number of users reduces to 4. For the community C_4 is interested in the topic "international cooperation", in Apr-2009, the number of users in C_6 is 30, but in May-2009, the community C_7 reduces to 21 users. Analyzing the topic "admission", we see the peak of community C_6 in Apr-2009 is 56. In 3 months May-2009, Jun-2009 and Jul-2009, there aren't any communities interested in "learning and examination" and "admission" topics. The community on topic "international cooperation" is relatively stable during the analysis period in figure 11 from Dec-2008 to Jul-2009. Thus, the elasticity of the number of users in communities indicates the phenomenon of joining or leaving the communities of users. That means at the point t_i there are more or fewer users in communities than the t_{i-1} or t_{i+1}.

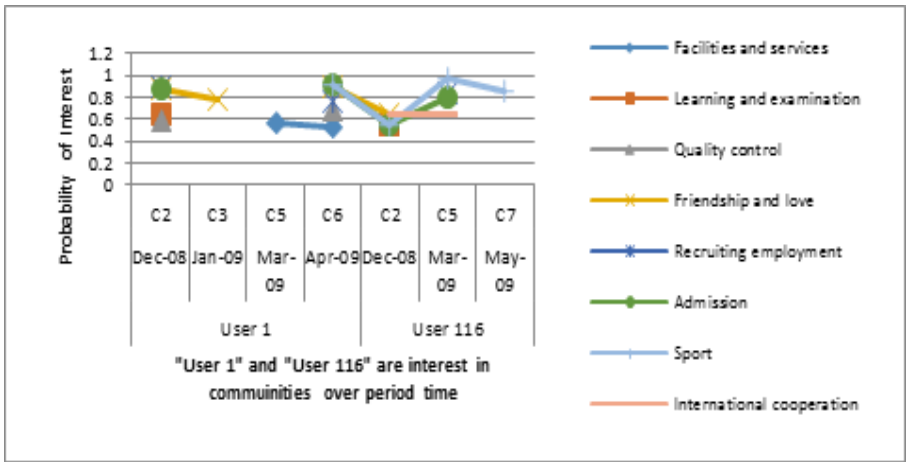

Fig. 12. The change interested topic of "user 1" and "user 116" of communities on topic in period from Dec-2008 to Jul-2009 by interested probability

Figure 12 shows the communities which "user 1" and "user 116" join in. We survey random with two users "user 1" and "user 116" on 9 topics. With user "user 1" joins in communities are C_2, C_3, C_5 and C_6 with each different interested probability on each different topic. And user "user 116" joins in communities are C_2, C_5 and C_7. These users have the changes about interested topic, probability and community in each period of time.

D. Results Evaluate

Application of the Precision (P), Recall (R) and F-measure (F) in [14] to evaluate the clustering results by Kohonen network. We compare the results of clustering vector of topics according to the proposed model and the clustering results by manual [19][23]. Assume that in set of actors we divide these actors into m clusters of actors by manual (by clustering based on the topics of the forum). On the other hand, by using SOM, the set is split to k clusters. Precision measure represents the ratio of the accuracy of a SOM cluster. If the ratio is 1, it means that all the actors in SOM cluster belong to cluster m_i, or $k_i \subset m_i$. Precision measure represents the ratio of the accuracy of a SOM cluster. If the ratio is 1, it means that all the actors in SOM cluster belong to cluster m_i, or $k_i \subset m_i$. According to Brew & Schulte im Walde (2002), F-Measure, which is the combination of Precision and Recall, is used to compute the accuracy of the system. For the clustering system, this is the equation:

$$F = \frac{2PR}{P + R} \tag{12}$$

The greater value F-Measure has, the more accurate the SOM is. Theo Brew C. [4] proposed evaluation method follows: corresponding to a cluster in the clustering result of the system we calculate the value of the F-measure for all clusters to be created manually. Choosing cluster which has the value of the highest F-measure and remove that cluster and repeating the above step for the remaining term. The total values of F-measure higher clustering system more accurately.

Here are the results of the corresponding F-measure (see table 3) with m = 5 clusters and k = 6 clusters (by Kohonen). We we compute the table of Precision, Recall, then manipulate the total F-measure.

TABLE III. THE RESULTS OF F-MEASURE VALUE BETWEEN MANUAL (CLUSTER BASED ON THE TOPICS OF THE FORUM) AND KOHONEN

Kohonen/Manual	**C_0**	**C_1**	**C_2**	**C_3**	**C_4**
C_0	0.43	0.15	**0.84**	0.52	0.68
C_1	0.67	0.61	0.00	0.16	0.00
C_2	0.00	0.36	0.51	**0.62**	0.16
C_3	0.72	0.00	0.55	0.55	0.34
C_4	**0.81**	**0.73**	0.25	0.00	**0.72**
C_5	0.19	0.00	0.15	0.29	0.36
MAX	**0.81**	**0.73**	**0.84**	**0.62**	**0.72**

Total MAX for clustering by Kohonen network is:

$$0.81 + 0.73 + 0.84 + 0.62 + 0.72 = 3.72.$$

Total max value of F-measuare in table 4 is 3.72 (respectively 74%). This value according to our assessment is high, this proves the proposed method using the clustering method of Kohonen network combined TART model with high accuracy.

VI. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

The contributions to this paper are summarized into two major issues:

1) Proposing a new model to discover online communities based on the topic model:

We focus on exploiting and combining Kohonen network and TART model. The model consists of two main components: (1) standardizing and selecting the result from the output of TART model. This is a set of interested topic vectors of users on social networks and is also a set of input vectors for Kohonen network, (2) proposing the model of using Kohonen network to discover communities of users interested in specific topics which are called communities of users on topics. The model can discover users' interested topics in each period of time and probability of topics interests, calculating topics apportion according to each online community. The challenge given in this content is to discover online communities through discussed contents because communities frequently change interested topics as well as members who participate in social network communities.

2) Analyzing changes in interested topics and users in communities on social networks in each period of time is based on the output layer of SOM and the relationship among that output layer.

B. Future work

The results of this paper will be the basis for researches in the future such as looking for important people in communities, analyzing the influence spread of topics and searching for the origin of information on social networks.

ACKNOWLEDGEMENT

This research is funded by Viet Nam National University HCM City (VNU-HCMC) under Grant number B2013-26-02.

REFERENCE

[1] Alexandru Berlea1, et al., Content and communication based sub-community detection using probabilistic topic models, IADIS International Conference Intelligent Systems and Agents, 2009.

[2] Andrew McCallum, Andr´es Corrada, Xuerui Wang, 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, 2004.

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

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

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

[6] D. Zhou et al., Probabilistic models for discovering e-communities. In WWW '06: Proceedings of the 15th international conference on World Wide Web, page 182. ACM, 2006, pp. 173-182.

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

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

[9] Ho Trung Thanh, Do Phuc, Ontology Vietnamese in Higher Education, Journal of Science and Technology, Vietnam Academy of Science and Technology, Volume 52, No. 1B, 2014, pp. 89-100.

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

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

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

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

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

[15] Kohonen, T. , *Self-Organizing Maps*. Extended edition. Springer.

[16] Kohonen, T. and Somervuo, P. (2002). How to make large self-organizing maps for nonvectorial data. *Neural Networks* 15(8-9), 2001, pp. 945-952.

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

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

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

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

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

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

[23] Thanh Ho, Phuc Do (2015), Analyzing Users' Interests with the Temporal Factor Based on Topic Modeling, 23-25 March 2015, Indonesia, Springer, 1982, pp. 106-115.

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

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

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

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

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

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

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