ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Detecting Communities and Surveying the Most Influence of Online Users

Thanh Tran[1], Thanh Ho[2] and Phuc Do[1]

[1]University of Information Technology, VNU-HCM, Vietnam
duythanhcse@gmail.com, phucdo@uit.edu.vn

[2]Faculty of Information System, University of Economics and Law
VNU-HCM, Vietnam
thanhht@uel.edu.vn

Abstract

Social network is a virtual environment that provides services for connecting users with the same interests, points of view, gender, space and time. Beside connection, information exchange, communication, entertainment and so on. Social network is also an environment for users who work in online business, advertisement or politics, criminal investigation. How to know what users discuss topics via exchanged contents and communities which users join in? In this paper, we propose a model by using topic model combined with K-means to detect communities of online users. Each user in social network is represented by a vector in which the components are the distribution probabilities of interested topics of that user. Based on the components of this vector, we discover the interested topics of online users to detect communities and survey users who are the most influence in communities to recommend for spreading information on social network.

Keywords: *LDA, ART, K-Means, online community, topic model, influence.*

1. Introduction

The social network has become a familiar concept of information technology. Positive or negative impacts are shown through the analysis of social networks, which is much more important than the work of capturing information and settling information in the real social life. Actually, at present there are a lot of issued researches on the social network analysis [1][2][3][4].

The social structure of social networks represented like a human society in the real life is known as the online community [4][5][6]. A social network is a heterogeneous huge data set, with many links represented by a graph. In the graph, with the actor corresponding to the object and the edge corresponding to the link in the interactive relationship between the objects, a social network will have the similarities in online communities, such as the similarity in friend relationships, the similarity in interests, and the similarity in affinities of other characteristics, including work, education, social interaction [1][2][3]. The social network for clustering is to find out the characteristics similar to online communities put into groups according to specific topics.

Figure 1. Communities on social networks[1]

Clustering social networks has many implications in management, economic activities, science and society. Social networks with lots of data to analyze at present have three main types of data which are often analyzed as follows:

Firstly, the analysis is based on the friend relationship: this analysis, which is mainly based on the relationship of friends in community on the social network, has important implications in identifying the strength in community relations [5][6][7]. This helps managers make decisions effectively in their organizations and we can also determine the characteristics of a community when we know a few online communities like the expression "please tell me who is your friend, I will tell you how you are ".

Secondly, the analysis is based on the exchanged contents [7]. This analysis helps understand that the interest swap of users happens on social networks. Clustering the relevant exchanged contents really helps us cluster

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

communities sharing their same opinions to strengthen collaboration.

Thirdly, clustering social network communities consists of structure and content [8]. This helps find out the online users in communities which have the same structure and content. From that, managers can easily communicate and create the most efficient group collaboration.

Based on the features of communities, we can find out communities users who are the most influence in communities.

Figure 2. Influential factor of community members[2]

In the paper, we propose a model for detecting interested topics via exchanged contents on the social network based on Latent Dirichlet Allocation model [9] and Author-Recipient-Topic [10], and then proceed detecting the community by algorithm K-Means [11] combined with the topic model.

The paper is organized as 1) Introduction 2) Related work 3) General model for detecting community based on K-means and topic moel 4) Experiments and discussions 5) Conclusion and future work.

2. Related work

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete text data [10]. In general, LDA is a three-level hierarchical Bayesian model [9][10][12] in which each document is described as a random mixture over a latent set of topics. Each topic is modeled as a discrete distribution of a set of words. LDA is suitable for the set of corpus and the set of grouped discrete data. LDA can be used for modeling the document on the purpose of detecting some underlying topics of that document. The generative process of a set of documents consists of three steps: (i) each document has a probabilistic distribution of its topics; this distribution is estimated as the Dirichlet distribution. (ii) for each word in a document, a specific topic based on the distribution of the topics of that document is chosen (iii) each keyword will be chosen from the multinomial distribution of the keywords according to the chosen topic [10][12].

The purpose of LDA is to detect each word belonging to a specific topic. From that we can guess the label of that topic. The importance of topic model is the posterior distribution. This can be seen as the generative process and the posterior inference for the latent set of variables, which are the keywords of the topic. In LDA, this process is calculated by the equation:

$$p(\theta, \emptyset, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \emptyset, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \quad (1)$$

In the equation (1), we have the variables z, θ, ϕ. For each θ_j which is a vector of topics of document j, z_i is the topic of word w_i, $\phi^{(k)}$ is the matrix KxV with $\phi_{i,j} = p(w_i \mid z_j)$

However, in equation (1), we can't precisely calculate with the normal factor $p(w \mid \alpha, \beta)$. Therefore, we normally use Gibbs Sampling (Griffiths & Steyvers, 2004; Steyvers et al., 2004; Rosen-Z vi et al., 2004) for inference.

2.2 Gibbs Sampling

Gibbs Sampling is a member of a family of algorithms from the Markov Chain Monte Carlo (MCMC) [13]. The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations through the chain, sampling from the distribution should converge to be close to sampling from the desired posterior. Gibbs Sampling [10][12][13][14] is based on sampling from conditional distributions of the variables of the posterior. For example, to sample x from the joint distribution $p(x) = p(x_1, x_2, \ldots, x_m)$. We do not have any proper solution to compute p(x), but a representation for the conditional distribution is possible, using Gibbs Sampling perform the following steps [12]:

1. Randomly initialize each x_i
2. For t=1…T:
   2.1. $x_1^{t+1} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots, x_m^{(t)})$
   2.2. $x_2^{t+1} \sim p(x_2 | x_1^{(t)}, x_3^{(t)}, \ldots, x_m^{(t)})$
   2.3. $x_m^{t+1} \sim p(x_m | x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-1}^{(t)})$

2.3 ART model (Author - Recipient - Topic)

ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent [5][10]. In its generative process for each message, an author a_d and a set

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

of recipients $\boldsymbol{r}_d$ will be observed. To generate each word, a recipient x is chosen from the set $\boldsymbol{r}_d$; and then, a topic z is chosen from a multinomial distribution $\theta_{a_d,x}$. This distribution is specified with the *author-recipient* pair (a_d, x). Finally, the word w is generated by choosing samples from a multinomial distribution $\emptyset_z$. The process of choosing samples is based on the *Gibbs sampling* algorithm. The final result is the discovery of topics in a social network where the messages are created.

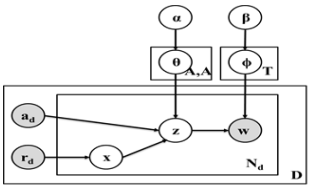

Figure 3. ART model [10]

In ART, given hyperparameters α and β, the author α_d and the set of recipients r_d,the joint distribution of an author mixture θ , a topic mixture $\emptyset$, a set of recipients x_d (belonging to X_d), a set of topics z_d (belonging to N_d), and a set of words w_d (belonging toN_d) is given by:

$$p(\theta, \emptyset, \boldsymbol{x}_d, \boldsymbol{z}_d, \boldsymbol{w}_d | \alpha, \beta, a_d, \boldsymbol{r}_d) \quad (2)$$
$$= p(\theta|\alpha)p(\emptyset|\beta)\prod_{n=1}^{N_d} p(x_{dn}|\boldsymbol{r}_d)p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\emptyset_{z_{dn}})$$

Integrating over θ and $\emptyset$, summing over $\boldsymbol{x}_d$ and $\boldsymbol{z}_d$, we get the marginal distribution of a document:

$$p(\boldsymbol{w}_d|\alpha, \beta, a_d, \boldsymbol{r}_d) \quad (3)$$
$$p(\theta|\alpha)p(\emptyset|\beta)$$
$$= \iint \prod_{n=1}^{N_d} \sum_{x_{dn}} \sum_{z_{dn}} p(x_{dn}|\boldsymbol{r}_d)p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\emptyset_{z_{dn}})d\emptyset d\theta$$

Finally, we have the probability of a corpus is:

$$p(D|\alpha, \beta, a, r) = \prod_{d=1}^{D} p(w_d|\alpha, \beta, a_d, r_d) \quad (4)$$

ART model describes the interaction of each node by analyzing transferring information of each node in the network; a topic relates to author, recipient and discovers role of author and recipient in transferring information process. Hence, the identification of topics in ART model depends on the social network in which messages are sent and received. Each pair of sender and receiver has a distribution over topics and each topic has a distribution over words.

2.4 Community-Author-Recipient-Topic model (CART)

In [15], the authors introduce CART model (Community - Author - Recipient - Topic), the model is tested on the Enron email data system [3]. 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 [15].

2.5 Finding the cluster of actors model

In [16], the authors present how to use SOM network to cluster the actor based on interested topic vector from dataset in English. This vector is a distribution probability of topic that actor prefers. The authors use ART model to create the vector of interested topics and use Enron email corpus as a sample dataset to evaluate efficiency in SOM network. By experimenting on the dataset, the authors demonstrate that our proposed model can be used to extract well and meaningful cluster following the topics [16].

2.6 Motivation research

We propose a model by using K-means algorithm combined with topic model for clustering online users based on their interested topics to detect online communities. These topics are exploited from a corpus of messages in Vietnamese on social networks via exchanged contents of users. Besides that, we survey the users who are the most influence in communities for spreading information on social network.

3. General model

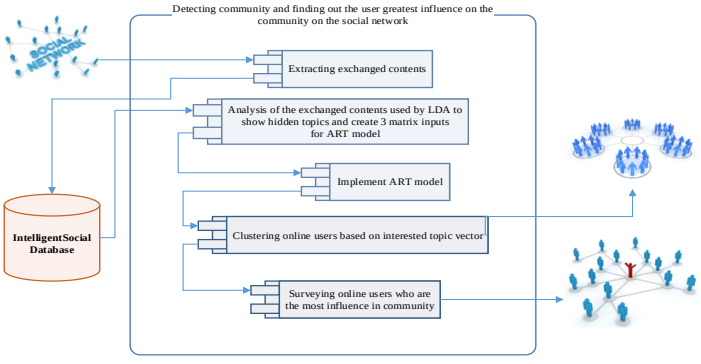

Figure 4: Model of detecting communities and finding out the users with the most influence in communities

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

We propose a model as shown in Figure 4. Our model consists of information extraction, data cleaning, social network analysis to find out the interested topics by using the LDA model, labeling topic and detecting online community. There are 5 steps, including:

Step 1: We do the data cleaning process for the social networks dataset. Each message will be characterized by keywords and removed the stop words.

Step 2: After cleaning the data and using the LDA model, we will have the matrix words of the topics T x V (word, the distribution probability) and the matrix distributed the messages based on the topics T x D (message's id, the distribution probability) .

Step 3: Applying the matrixes T x V and T x D for ART model to create interested topic vectors. Each online user has a interested topic vector. Each vector of online users based on interested topics is a vector representing the interested probability of the topics of each user in a social network. Each user can have one or many interested topics.

Step 4: Using K-means algorithm for clustering of users on social networking based on interested topic vector created in step 3.

Step 5: Surveying users who are the most influence in communities to recommend for spreading information on social network.

4. Experiment and discussion

4.1 Input dataset

The dataset is collected from Facebook, include:

- 75740 posts and comments in Vietnamese
- 2315 online users.
- 10 topics are surveyed.
- From 2014 – 2015 (2 years).

4.2 Implementation

After cleaning the data, by using the LDA model with the parameters $\alpha = 0.5$, $\beta = 0.1$, the number of iterations for Gibbs sampling is 2000, the number of steps is 100, the number of topics is 50 [10][12][14]. We have the matrix words of the topics T x V (word, the distribution probability).

After using the LDA model to figure out the topics, we use K-means algorithm [11] to group the messages into cluster. We consider each cluster as community. There are *n* users, each user has *m* attributes, and we divide them into *k* communities based on their attributes by using the K-means algorithm.

For clustering exchanged contents: we analyze the exchanged contents of users on social networks to find out interested topics vector, and then use the K-Means algorithm to cluster community based-on interested topics vector of users. In a community includes users who have the same interested topics. To do this, we study and implement the model ART (author – recipient – topic).

During the analyzing process, the ART model will create three matrices: the distribution matrix of words according to the distribution matrix of topics, authors, recipients and messages.

a. Matrix 1: The message - author – recipient (table 1)

This matrix contains message - author - recipient, each line consists of message code, author code and recipient code. To get this matrix, we extract and analyze information from tables of posts, comments and profile of users in the dataset.

Table 1: Matrix message - author - recipient

Message (ID)	Author (ID)	Recipient (ID)
629	200	196
630	200	196
631	200	196
632	222	196
633	219	196
634	222	196
635	222	196

b. Matrix 2: message - vocabulary - frequency of appearance

Table 2 – Matrix message - vocabulary - frequency of appearance

Message (ID)	Vocabulary (ID)	Frequency of appearance
633	220	1
633	20636	1
633	65	3
633	5343	1
634	15084	1
634	16273	1

Each line of this matrix contains the message ID, vocabulary ID and frequency of appearance in the message. This matrix usually contains volume of data very large because it must analyze each message contained in Matrix 1, for each message must parse out the matrix 2.

c. Matrix 3: Vocabulary – Vocabulary ID

Building more than 23,000 words extracted from the VnTokenizer tool, this is a project under the state was announced.

Table 3 - Matrix vocabulary – vocabulary ID

Vocabulary	Vocabulary (ID)
khử_trùng	5314
vội_vã	6038
Josu	6970
nhân_quyền	16488
bản_thân	14221
nhoè_nhẹt	8062
Hai_Hoàng	17589
Dương_Minh_Quang	6408
Tanimex	6971
Lê_Dũng	8265
ngập_chìm	17350

4.3 K-means algorithm combined with topic model for clustering communities

We use the K-means with the parameters: 10 topics, the number of messages is 75740, 2315 users, the number of topics is 10, the number of communities is 4 (k=4). We have experimented with k = 2, k = 3, k = 4, k = 5, k = 6,.., k = 10 ... to have a comparable clustering results and choose k = 4.

Figure 5 shows in detail the number of users (actors) belonging to the cluster (4 clusters):

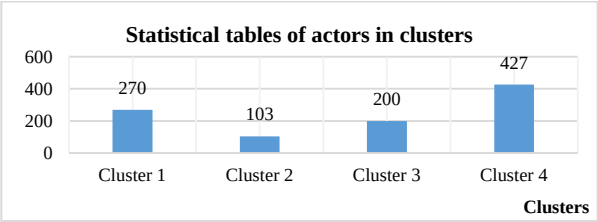

Figure 5 - The number of users in each cluster and the number of clusters. Cluster 1 has 270 users, cluster 2 has 103 user, cluster 3 has 200 users and cluster 4 has 427 users.

Table 4 – The set of vector centroid of 4 clusters

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
C1	0.00	0.00	0.00	0.28	0.01	0.02	0.02	0.43	0.00	0.24
C2	0.00	0.11	0.16	0.05	0.04	0.01	0.19	0.00	0.34	0.10
C3	0.49	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.20
C4	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

From the data in table 4, we easily know how much each cluster interested topic vectors and Distribution the topics according to vector centroid of 4 clusters (communities).

Figure 6 shows 4 communities (C1, C2, C3 and C4), each community has topics. For example, community C1 has 6 topics (T3, T4, T5, T6, T7 and T9), C2 has 8 topics (T1, T2, T3, T4, T5, T6, T8 and T9), C3 has 4 topics (T1, T2, T8 and T9) and C4 has 2 topics (T1 and T2). In which, C2 has the number of topics more than C1, C3 and C4.

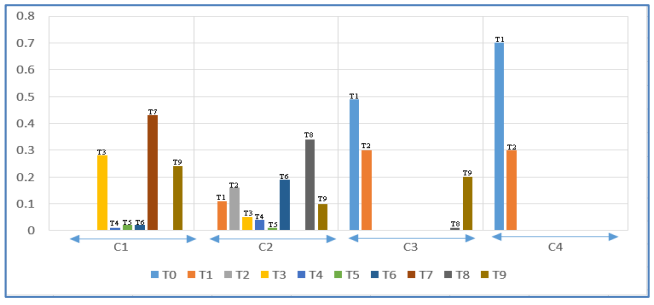

Figure 6 - Distribution 10 topics (from T1 to T9) according to vector centroid of 4 clusters (communities)

Figure 7 and figure 8 show result of detecting communities. There are 4 communities related to 4 clusters.

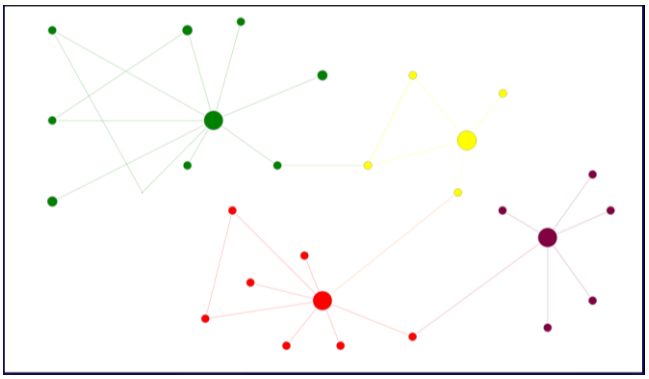

Figure 7. Results of detecting communities without user's name on nodes and surveying user who has the most influence in communities

.

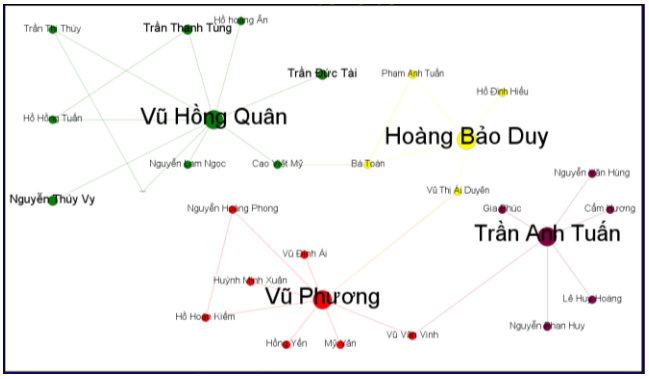

Figure 8. Results of detecting communities and surveying user who has the most influence in community

Each community was distinguished by a different color, if the user has a relationship of other users, they will be linked together. In the illustration above, we have 4 clusters with 4 users who are the most influence in communities such as: "Vũ Hồng Quân" in community 1, "Hoàng Bảo Duy" in community 2, "Vũ Phương" in community 3 and "Trần Anh Tuấn" in community 4 (see figure 8).

We can see the result {vuhong.quan.73, tough.crystal, nhung.vu.58760, ...} of cluster 1, {hoangbaoduy, transleyhan, nguyenhieu08 ...} of cluster 2 and so on (see table 5).

5. Conclusions and future work

5.1 Conclusions

Our research has proposed propagation models and algorithms through the analysis of social networks based on the specific topics. Especially, the research focuses on finding the topics with LDA model and clustering exchanged contents by using K-means algorithm combined with the topic model.

- Building the automated tools to retrieve exchanged contents from Facebook: All the exchanged contents consist of online users, posts, comments, likes, etc.
- Experimenting the detection topic module with LDA model, will help select a number of interested topics, such as political security, science and technology, sports, culture, arts, health and education to carry out clustering community on social networks.
- Experimenting Author - Recipients - Topics (ART) model will help create the set of vectors with interested topics of online users to provide community clustering.
- Proposing the community clustering model by using K-means algorithm combined with the topic model is to cluster online users based on interested topics vectors to find out online communities. After detecting community, we survey the user who has the most influence in community in order to recommend for spreading information on social networks.

5.2 Future work

In the future, we will continue to study and give recommendations in order to evaluate the results of the cluster as well as the quality of the proposed model.
Clustering the exchanged contents of online users helps us find out communities on social networks. In this community, there will be the same interested topics of online users who will have no relationship with other online users. Therefore, in order to cerate the relationship between the online users having the same interested topic with other online users, we will propose a model to spread the information to other online users on social network based on the users who are the most influence in community.

Acknowledgments

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-26-02.

Table 5 - Matrix probability of interest topic vectors of online users

Topic / Online users	T-0	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9
vuhong.quan.73	0.00600	0.00110	0.0031	0.2000	0.05000	0.1000	0.1008	0.350	0.000	0.189
tough.crystal	0.00601	0.00105	0.0030	0.2020	0.05194	0.1000	0.0990	0.347	0.000	0.190
nhung.vu.58760	0.00700	0.00120	0.0030	0.2010	0.05000	0.1010	0.0868	0.350	0.000	0.200
motcoidive.hue	0.00150	0.00000	0.0000	0.2985	0.00000	0.0000	0.0000	0.450	0.000	0.250
hoangbaoduy	0.00000	0.00200	0.0000	0.0000	0.00100	0.0000	0.4970	0.000	0.300	0.200
transleyzhan	0.00000	0.00000	0.5801	0.0000	0.00000	0.0209	0.0000	0.000	0.399	0.000
nguyenhieu08	0.00000	0.00000	0.5799	0.0000	0.00000	0.0201	0.0000	0.000	0.400	0.000
vu.phuong.5264	0.50000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
nguyen.vietanh.338	0.51000	0.29000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
diem.dtk	0.50000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
Tuan777	0.70000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.000
truclieu.nguyen.3	0.70000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.000

References

[1] J.Leskovec, L.A.Adamic, and B.A.Huberman, (2007). "The dynamics of viral marketing". In ACM Trans, volume 1.
[2] Muon Nguyen, Thanh Ho, Phuc Do (2013), *Social Networks Analysis Based on Topic Modeling*, The 10th IEEE RIVF International Conference on Computing and Communication Technologies (P.119-122), Hanoi.
[3] P.Domingos and M.Richardson, (2001), "Mining the network value of customers". In Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 01, pages 57–66, New York, NY, USA.
[4] Mr. Sachan, D. Contractor, T.A. Faruquie, L. V.Subramaniam, (2009), Using Content and Interactions for Discovering Communities in Social Networks. April 16–20, 2012, Lyon, France.
[5] Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang (2007), Topic and Role Discovery in Social Networks, Journal of Articial Intelligence Research 29.
[6] M. Sachan, D. Contractor, T. A. Faruquie, and L. V.Subramaniam. Probabilistic model for discovering topic based communities in social networks. 2011.

[7] Chunshan Li, William K. Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, Xin Li (2014), The Author-Topic-Community model for author interest profiling and community discovery, Springer-Verlag London 2014, pp. 74-85.
[8] The Anh Dang, Emmanuel Viennet (2012), Community Detection based on Structural and Attribute Similarities, ICDS 2012 : The Sixth International Conference on Digital Society, pp. 7-14.
[9] David M.Blei, (2003), "Latent Dirichlet Allocation". Computer Science Division, University of California, Berkeley, CA.
[10] Andrew McCallum, Andr´es Corrada, Xuerui Wang, (2004), The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email, Department of Computer Science, University of MA.
[11] http://en.wikipedia.org/wiki/K-means_clustering
[12] Tom Griffiths (2004), Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation - Gruffydd@psych.stanford.edu.
[13] B. Walsh, (2004), "Markov Chain Monte Carlo and Gibbs Sampling". *Lecture Notes for EEB 581, version 26 April 2004.*
[14] William M. Darling, (2011), "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling". *School of Computer Science University of Guelph.*
[15] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson (2008), Social topic models for community extraction. The 2nd SNA-KDD Workshop, vol 8.
[16] Tran Quang Hoa, Vo Ho Tien Hung, Nguyen Le Hoang, Ho Trung Thanh, Do Phuc (2014), Finding the Cluster of Actors in Social Network based on the Topic of Messages, ACIIDS 04/2014, ThaiLan. Springer, pp. 183-190.

First Author. MS. Thanh Tran got master's degree at University of Information Technology, VNU-HCM, Vietnam. His strong ability is about website, mobile and desktop applications with the programming languages as Java, Php, Objective C, Android, and the database as MySQL and Oracle.

Second Author. MS. PhD Student. ThanhHo works for Faculty of Information System, University of Economics and Law, VNU-HCM, Vietnam. His interests are data mining, e-commerce, Business Intelligent, social network analysis and management information systems. He is a member of Prof. Do Phuc's project.

Third Author. Prof. Do Phuc works for the University of Information Technology, VNU-HCM, Vietnam. His interests are data mining, bioinformatics and social media analysis. His current project is toward the analysis of social network based on thecontent and structure.