Discovering Hot Topics On Social Network Based On Improving The Aging Theory

Thanh Ho[1], Duy Doan[2], Phuc Do[2]

[1]**University of Economics and Law, VNU-HCM, VietNam**
thanhht@uel.edu.vn

[2]**University of Information Technology, VNU-HCM, VietNam**
dvnduy@gmail.com, phucdo@uit.edu.vn

Abstract

In this paper, we focus on discovering information on the social network to discover automatically the hot topics in real time. We suggest using aging theory for finding hot topics and using the tools for processing natural language and data processing model. We create a perfect general model. This system will help us find hot interesting topics on the social network in real time. With this idea, we will build a survey system which is the most accurate overview of the real world. If we can discover the interesting topics on the social networks, we can solve many problems in our present lives such as: to predict favored industry, the rising of consumer goods, Internet user community's psychology, the objects that exchange information for each event, and to determine the direction of the social network.

Keywords: *Social network analysis, LDA model, SVM model, aging theory*

1. Introduction

The evolving and high bursting of the social network bring benefits to the users. This virtual world helps us connect with each other. It helps us exchange information through messaging, posting, and other utilities. We usually surf the net every day, and we catch the news or the articles that have the same content from some sources on the social network, news from the internet or the forums. Most people are attentive with this information, and we wonder why it makes them curious and how many problems are they interested in. How do these problems affect people? We assume that there is a system that will help us discover information and predict which information is interesting. With this, it will help us make decisions and understand the situation in our society. As a result of this, we came up with the research "Discover the hot topics on the social network."

With this idea, we will build a survey system which is the most accurate overview of the real world. If we can discover the interesting topics on the social networks, we can solve many problems in our present lives such as: to predict favored industry, the rising of consumer goods, Internet user community's psychology, the objects that exchange information for each event, and to determine the direction of the social network. The rest of this paper is organized as follows: 2) Related works, 3) Hot topic discovery model, 4) Discover hot topic, 5) The test results and discussion, 6) Conclusions and future works.

2. Related Works

2.1 Word segmentation

vnTokenizer. This tool is developed in VLSP research [8] of Le Hong Phuong, its accuracy is 97%.
jvnTagger. It is a tool to determine the type of word base on Conditional Random Fields (Lafferty et al., 2001) and Maximum Entropy (Nigam et al., 1999). JVNTagger is built by the state-level topic VLSP [8] with training data of 10.000 sentences and 20,000 sentences of Viet Tree Bank. It was tested by 5-fold cross validation method on VTB-10,000. The result of CRFs can get F1's biggest value which is 93.45% and 10-fold cross validation with Maxent/VTB-20.000 can get F1' highest value of 93.32%.

2.2 Improve word segmentation

We realized that the word segmentation method process very slowly in the actual test, so we recommend two methods to improve processing speed when we use this tool for word segmentation. With this way, we reduced error rates 55-folds and increased 2-folds processing speed

Improve Speed. Actually when we use this tool for text analysis, we realized that it was limited by processing speed, We should divide the text into many sentences before we start to process the text by vnTokenizer [8]. We will combine these sentences to complete the text when the system finishes the processing for each sentence by vnTokenizer. This way will improve the speed when we process it with big data. In fact, it is six times faster than the old way

Improve quality for text. jvnTager is used to determine the type of word. It has also many errors when we use it. For example: it cannot determine wrong vocabulary, wrong grammar, special symbols, html tags, stop words, etc. When this tool usually contains the errors when it meets the above problems, we use data pre-processing method using the automata theory. In this research, we use "regular expression tool". It is a type of automata language. This tool's purpose is to clean data before we process the text through jvnTagger [8]. With this way, we reduced error rates 2-folds based on 13,208 texts

Table 1. Compare results between old method and new method

2000 texts for each phase	*Number of the errors with old method*	*Number of the errors with new method*
phase 1	120	11
phase 2	50	0
phase 3	280	13

2.3 LDA algorithm and Dirichlet Distributions

In the LDA model [1][6][9], we would like to say that the topic mixture proportions for each document are drawn from some distribution, so we want to put a distribution on multinomial. That is, k-tuples of non-negative numbers that sum to one. The space is of all of these multinomial has a nice geometric interpretation as a (k-1)-simplex, which is just a generalization of a triangle to (k-1) dimensions. Criteria for Dirichlet Distributions

- It needs to be defined for a (k-1)-simplex
- Algebraically speaking, we need a multinomial distribution (it is a probability distribution for a multivariate discrete random variable)

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \qquad (1)$$

Useful Facts.

- This distribution is defined over a (k-1)-simplex. That is, it takes k non-negative arguments which sum to one. Consequently it is a natural distribution to use over multinomial distributions.
- In fact, the Dirichlet distribution is the conjugate prior to the multinomial distribution. (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet)
- The Dirichlet parameter α_i can be thought of as a prior count of the i^{th} class

Gibbs Sampler. in its basic incarnation, is a special case of the Metropolis Hastings algorithm [5]. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Suppose we want to obtain k samples of $X = (x1, \dots, x_n)$ from a joint distribution $p(x_1, \dots, x_n)$. Denote the i^{th} sample by $X^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$. We process as follow

- We begin with some initial value $X^{(0)}$
- For each example $i \in \{1 \dots k\}$ sample each variable $x_j^{(i)}$ from the conditional distribution $p\left(x_j | x_1^{(i)}, \dots, x_{j-1}^{(i)}\right), x_{j+1}^{(i-1)}, \dots, x_n^{(i-1)}$ That is, sample each variable from the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha} \qquad (2)$$

2.4 SVM algorithm

Support Vector Machines (SVM) [3][7][10][11] is a type of classification algorithm based on determining the optimal division between two sets of feature vectors. In the simplest form of SVM, this division is linear, while an extended form of SVM utilizing a "kernel function" allows non-linear classification. One characteristic of SVM classifiers is that they can operate efficiently on data with large feature sets, otherwise described as data with high dimensionality. This is useful for many pattern recognition tasks; in particular, musical applications based on spectral features tend to have high degrees of dimensionality

Transductive support vector machines. Transductive support vector machines [7] extend SVMs in that they could also treat partially labeled data in semi-supervised learning by following the principles of transduction. Here, in addition to the training set D, the learner is also given a set

$$D^* = \{x_i^* | x_i^* \in R^p\}_{i=1}^{k} \qquad (3)$$

Of test examples to be classified. Formally, a transductive support vector machine is defined by the following primal optimization problem: Minimize $(in\ w, b, y^*)$

$$\frac{1}{2}||w||^2$$

subject to (for any i = 1, …, n and any j = 1,…, k)

$$y_i(w.x_i - b) \geq 1, y_j^*\left(w.x_j^* - b\right) \geq 1, \text{and}$$

$$y_j^* \in \{-1,1\} \qquad (4)$$

Transductive support vector machines were introduced by Vladimir N. Vapnik in 1998

3. Hot topic discovery model

Hot topic discovery model is a general model combined from tools, data processing method (data filter, vnTokenizor, jvnTagger), Latent Dirichlet Allocation model, support vector machine model and hot topic discover method based on the aging theory. This combination creates a general system that can solve our research problems. This model is a core system to discover the hot topics on the social network as shown in Fig. 1.

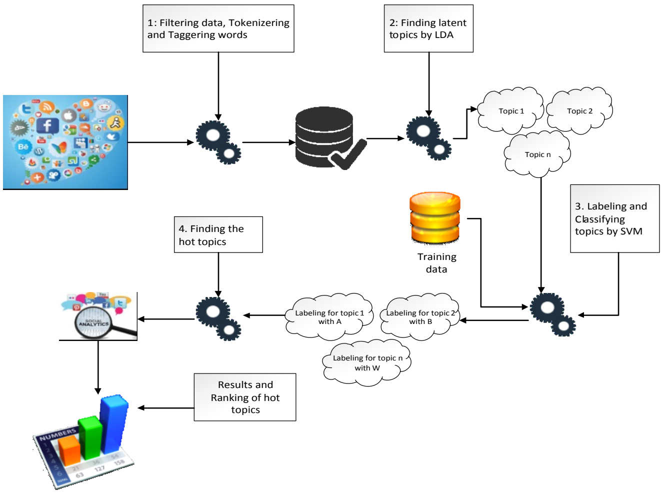
1: Filtering data, Tokenizering and Taggering words
2: Finding latent topics by LDA
Topic 1
Topic 2
Topic n
3. Labeling and Classifying topics by SVM
Training data
4. Finding the hot topics
Labeling for topic 1 with A
Labeling for topic 2 with B
Labeling for topic n with W
SOCIAL ANALYTICS
Results and Ranking of hot topics
NUMBERS

Fig. 1. General Model for hot topic discovery on the social network.

Stage one. The system crawls and integrates data from the forums. We need to pre-process data, separate types of words and set the label for words because the exchange information on the social network always has soiled data such as special symbols, shortcut words, the local language, mistaken vocabulary, incorrect grammar, stop words, etc. Therefore, the data filter is very important. The system will filter out grimy information or garbage data out of the text. Then we will will separate the words by some methods such as vnTokenizer [8], the method used to determine which is the single word and which is the compound word in Vietnamese. Lastly, we will use vnTagger tool [8] to determine the type of word, which is noun, adjective or verb. When the system determines the type of word, it will help us epitomize the content and get only the word that is meaningful to use for the following stage.

Stage two. After the system has cleaned the data, this step will separate the topics into groups when we have exchange information from users on the social network. The similar contents will be coupled to a group that has the same topic. In this step, we use the LDA algorithm [1][6] to detect the hidden topics based on the number of conversations on the social network. However, these topics don't have titles yet. It has only general name such as topic 1, topic 2, topic 3 and so on. The LDA algorithm will separate number of suitable conversations to the hidden topics.

Stage three. The above result is a collection of topics that don't have label (the topic that doesn't have the title) in this step. We will know the title for each topic, but first we must build an ontology manually about the topics and the perfect training data for each topic. Secondly, we use SVM algorithm [7[[11], the above training data and the test data (we have this result in stage two) to separate the hidden topics and set title for them.

Stage four. When the third stage is finished, we have discover each topic, and these topics also have suitable titles. This step's permission is to detect the interesting topic by real time. This way is based on the aging theory [2][4] and calculating method for the topic's energy (we will present this research at section 4). As a result, we will discover the hot topics on the social network

4. Discover hot topic

4.1 Introduction

Although daily posts involve a bunch of topics, more than 60% of these posts just focus on few subjects. Hence, it is very meaningful to discover hot topics [4] on the social network. A hot topic includes the following.

Massive Posts. Only an attractive topic can assemble lots of user's discussion, which, in turn, becomes a prerequisite for a hot topic. This factor is comprised in our energy calculation process. Each post contains certain nutrition which can be transmuted into energy, therefore, subjects with more posts could gain more energy.

High Quality Posts. Compared with those junk posts (like "I agree", "good"), a hot topic always has more posts of high quality. The relationship among posts could help us to identify which posts have high quality.

High Cohesion. Since scattered content has less attraction, the content of a hot topic is usually compact and centralized. We use the threshold of Single-Pass clustering to strictly control the number of threads to form the topic.

Bursting. For a hot topic, it often gathers a large number of posts in a short period of time, and then gets to a stable state until it slowly disappears, which implies a life cycle of the topic.

4.2 The aging theory and hot topic discovery method

Based on the aging theory of biology, Chen [2] is the first person who used aging theory for the topic model. Each topic has a life cycle and a life span. Sometime it evolves quickly, but sometime it becomes slow, the slowdown can be seen as a subject which is not interesting or is not discussed. We can divide life cycle of the topic to four phases: birth, old age, sickness, and death. The aging theory used the concepts about energy to chase an interesting topic. Energy value is a performance for the degree of a hot topic. The topic's hotness level is increased when it becomes popular. Many people discuss it, however, the topic's energy will be decreased by time. In this paper, we use the concept about nutrition for each post and it can be seen as a material for a topic. Each post contains a specific nutrition value. It is called nutrients. Moreover, these nutrients will grow into energy through metabolism. Time is divide into each slot. At each slot, we use the nutrients to convert them to energy. This way helps us update energy value for the topic that we are surveying at each slot by real time.

4.3 Our Improvements recommendation

Improve the quality of posts. Following Donghui Zheng and Fang Li's researching [4] the post's quality is determined by degrees of similarity in the text's content (it is called alpha) without including other problems, for example: How many people like this post? What is the author's confidence level in this topic? These problems are very interesting in the social network. Nowadays, for example the posts on Facebook which have high quality will be noticed by most people. These people will click "like" for a post which they favor. Otherwise, if the author of the topic is a famous person or an expert then the quality of the topic is more valuable. Based on the number of people who like posts, the confidence level of the author and the degrees of similarity between the content of the post and the subject. The equation (5) below is a general formula to determine the value of the post.

$$N_p = T + (L * 0.1) + C_{user} \tag{5}$$

Where.

- User's confidence level $C_{user} = \frac{P_{user}}{\sum P_{AllUser}}$
- P_{user} the number of post that the user write in a topic
- $\sum P_{AllUser}$ the number of post that the user write in all topics
- L: number of likes
- T: Theta, it is the degrees of similarity between content of post and subject
  $T = \begin{cases} T, T \geq 0.5 \\ 0, T < 0.5 \end{cases}$

Improve energy value. We know that the bursting topic included a lot of exchanging information in a short time. Our purpose is to determine more objectives to inflect to the hot topic, so when we record this value, it will help us know exactly the time when the topic bursts. The energy value is calculated by formula (6) as follow.

$$E_t = \sum(N \times p) + brusting_t \tag{6}$$

Where:

- E is energy value for the topic at time t, the difference time then the value also is differed.
- N is nutrition value for each post in same a topic.
- P is the coefficient for each topic.
- $Brusting_t$: is bursting value for topic "t". $brusting_t = \frac{count(posts)}{p}$

Improve hot topic discovery method. Following Donghui Zheng and Fang Li's researching [4] the topic's energy value is determined on the last time of the survey. We realized this way is subjective and is not purely right in many cases because the energy's value always change by time. For example, the topic has three times of the highest energy value but only has one time of low energy at the last time, so we can't review the topic based on the last time, so we recommend an improvement. The topic has the highest energy value for many times that makes it the interesting topic. The code below is used for discovering hot topics on the social network

Algorithm 1. Discovering hot topic on the social network

```
numberSlot  = determineSlot(t1,t2,distance);
numberHour = determineHour(t1,t2);
ListTE ← NULL;
For each Topic t ∈ ListTopic        do
    E[0] ← point(numberHour);
    ListE ← NULL;
        For each slot s ∈ numberSlot   do
            N[s] ← getNutrition(s);
            E[s] ← E[s-1] + $\sum(N_s \times p) + Brusting_s - distance$;
            ListE.Add(E[s],s);
    ListTE.Add(ListE,t);
    ListTopHotTopic ← FindHotTopic(ListTE,TopN);
```

5. The test results and discussion

The exchanging information on the social network contains important data for analytic and processing in this research. The data is crawled from resources such as the economy forum, the health forum, the information technology forum, Facebook, Twitter, etc. In this research, we use the database of the student forum. This data is seen as a virtual social network that includes members in university. The statistics fact is for this database as database size: 350 Mb, number of the posts: 13208, number of the subjects: 3855, number of the users: 14190, number of the ban users: 201, number of the groups: 13, number of the likes: 232, number of searching: 2401, number of the administrators: 7. We apply the hot topic discovery model for this database. In the test processing, we have the results as follow

Table 2. This is the result between BBS's system and our new system. It is tested based on 2000 first posts

Topics	***BSS's system***	***Our new system***
Associations	0	78
International cooperation	0	51
Learning	0	0

Originally, we started with 2000 posts. For surveying, we created three new topics which are associations, international cooperation and learning. These topics did not belong to the old system. We used the LDA algorithm and SVM algorithm to detect conversations that these topics could have. With above results, we had 78 conversations referring to the associations, 51 concerning to the international cooperation and no conversations referring to the learning. We found that the new system could detect conversations of the new topics that the old system couldn't. Moreover, we could build any training data when we needed. Therefore, if we wanted to create a new survey for a new topic, we only created the training data for that topic and started to survey. Obviously, our results depended on the training data very much, so it could be seen as the gold data. We must really notice when creating training data. If we have clear and perfect training data, the result will be very good.

Table 3. This is result of classify data method with the BSS's topics; it is tested on 13208 posts.

BSS's topics	*BSS's system*	*Our new system*
Internet & Public Libraries	528	380
School Safety Improvements	312	218
Biographical Details	136	218
Economics and Business Studies	111	69
Social Sciences	103	72
Physical Education	273	112
Associations	0	520
International cooperation	0	350
Learning	0	153

In this case, we started with 13208 posts (all conversations in this database). Although, we had 3855 topics in the old system, we only showed results of some topics in a paper because 3855 topics were too long to present here. For example, with the topic about Internet & Public Libraries, we had 528 conversations in the old system, but in fact, we only had 380 conversations in the new system. Similar to the School Safety Improvements, the old system had 312 conversations, the new system had 218 conversations only. Because we used the training data to detect these topics in the new system, we only filtered conversations that had similar contents. Thus, the number of conversations in our result was always less than the old system but more exact. Finally, we saw that the topics on the old system contained the noise data more than our system. With this solution, we could filter the topics that only contained valuable conversations.

In Fig. 2 and Fig. 3, all data were published from January 03, 2008 to December 30, 2010. There were up to 3,855 threads totally including 13,208 posts. In average, there were 35 threads including 142.3 posts per day. All posts have been already preprocessed.

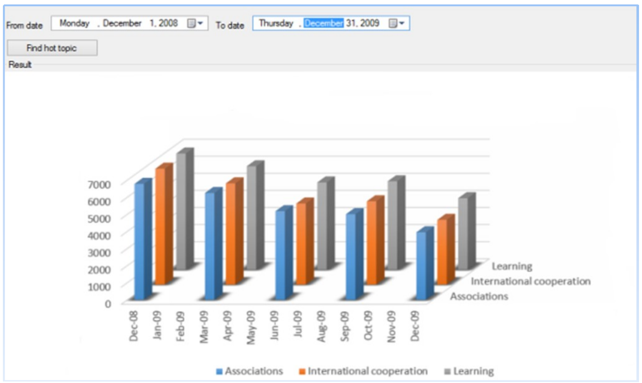
From date Monday , December 1, 2008 To date Thursday , December 31, 2009

Fig. 2. The topic's energy value by real time

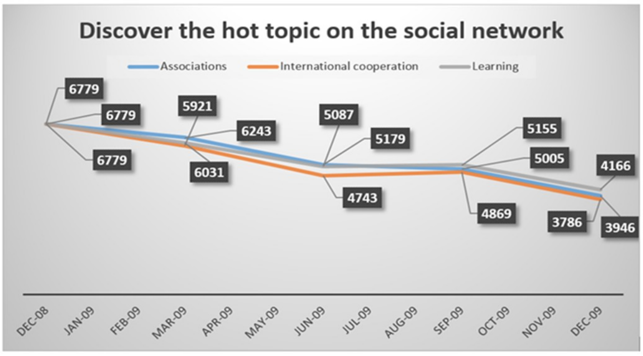

Fig. 3. The energy value distribution of the hot topics on the new system during from 12/2008 to 12/2009

In Fig.2 and Fig.3, we implemented a survey for three topics: the associations, the international cooperation and the learning. Our purpose was to find the hottest topic from December 2008 to December 2009. This time, we divided to 12 slots and we saw that the topic of associations had 8 times to be chosen as the hottest topic, from December 2008 to August 2009. Despite being after August 2009, the topic of associations was less hot than the topic of the learning, it had the most number of energy value more than others did. Therefore, the hottest topic from December 2008 to December 2009 was the associations.

With this test result, we consider that the number of conversations for each topic is not true in real life because our system detects the number of conversations for each topic is different from BBS's system. We can detect the hidden topic and filter noise conversations out of texts, so our system has a better result than BBS's system. Otherwise, we can create a new topic for surveying. We can know the number of conversations that topic can contain, after then the system will discover and report to us the hot topic by real time.

6. Conclusions and Future Works

Conclusions. This research's result combines the data mining models and the natural language processing tools. It brings benefits for data processing on the social network. This research also talks about investment method for hot topic discovery based on the aging theory and the investing method for natural language processing. This research brings optimistic results. The system can clean data automatically and filter data well. It also improves the hot topic discovery processing, classifying the topic, and setting label for each topic. The system's processing time will be decreased. It is an improvement.

Future works. We can expand research about data cleaning. It will increase the degree of accuracy for the data classification and setting label for the topic. We can build the data warehouse to integrate other resources such as Facebook, Twitter, Zingme, the e-news portal, the forum, etc. We also can integrate the Microsoft's technology platform about business intelligence (BI), SharePoint 2013 BI, SQL analysis services, SQL reporting service. With this, we will set up a model that can combine Microsoft's BI to analytic data. It will be a power system for the business intelligence system and the social network.

Acknowledgments

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

References

[1] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I, *Latent Dirichlet Allocation,* Journal of Machine Learning, vol. 3, pp. 993-1022 (2003)

[2] Chen, C.C, Chen, Y.T, Sun, Y, Chen, M.C: *Life Cycle Modeling of News Events Using Aging Theory* In: Lavrac, N, Gamberger, D, Todorovski, L, Blockeel, H.(eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 47–59. Springer, Heidelberg (2003)

[3] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer (2007)

[4] Donghui Zheng and Fang Li, *Hot Topic Detection on BBS Using Aging Theory* (2013)

[5] http://en.wikipedia.org/wiki/Gibbs_sampling

[6] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[7] http://en.wikipedia.org/wiki/Support_vector_machine

[8] http://vlsp.vietlp.org:8080/demo/?page=resources&lang=en

[9] T.Hofmann, *Probabilistic latent semantic analysis,* Uncertainty in Artificial Intelligence (UAI) (1999)

[10] T.Joachims, *Transductive learning via spectral graph partitioning*. Proceeding of The Twentieth International Conference on Machine Learning (ICML): 290-297 (2003)

[11] T.Joachims, *Transductive Inference for Text Classification using Support Vector Machines*. International Conference on Machine Learning (ICML) (1999)

First Author. MS. PhD Student. Thanh Ho 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.

Second Author. MS. Duy Doan works for the University of Information Technology, VNU-HCM, VietNam. His interests are Business Intelligent, social network analysis. 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 the content and structure.