International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.169-180
http://dx.doi.org/10.14257/ijdta.2016.9.5.17

Topic Discovery Algorithm Based on Mutual Information and Label Clustering under Dynamic Social Networks

Lin Cui[1,2], Dechang Pi[1] and Caiyin Wang[2]

[1]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing*
210016, Jiangsu, China
[2]*Intelligent Information Processing Laboratory, Suzhou University, Suzhou*
234000, Anhui, China
jsjxcuilin@nuaa.edu.cn

Abstract

In recent years, topic detection has become a hot research point of the social network, which can be very good to find the key factors from the massive information and thus discover the topics. The traditional label propagation-based topic discovery algorithm (LPA) is widely concerned because of its approximate linear time complexity and there is no need to define the target function. However, LPA algorithm has the uncertainty and the randomness, which affects the accuracy and the stability of the topic discovery. In this paper, a method for clustering label words based on mutual information analysis is presented to find the current topic. Firstly, through filtering the stop words and extracting keywords with TF-IDF, topic words are been extracted out, and then a common word matrix is built, a topic discovery algorithm based on mutual information and label clustering is put forward. Finally, extensive experiments on two real datasets validate the effectiveness of the proposed MI-LC (Mutual information-Label clustering) algorithm against other well-established methods LPA and LDA in terms of running time, NMI value and perplexity value.

Keywords: *Dynamic social network, Topic discovery, Mutual information, Label clustering*

1. Introduction

In recent years, social networks have become an important platform for people to exchange and share information, and their mode with the fission information sharing and communicating promotes the users more quickly to gather around the common concerned topics. Social network is a relatively stable relationship between individual members because of the interaction. Social networks are often abstracted into a graph, in which each person is represented with a node vertex and the relationships between people are denoted with the edges. Topic discovery in the social networks is of great significance for social network analysis.

With the research on the static properties about the social networks, researchers begin to pay attention to the time features of the social networks [1]. How to investigate the influence of time on the social networks and the changes of social networks with time becomes the new research direction of the social network. Traditional online social network topic detection is mainly based on the text topic extraction, lack of mining the topic dynamic evolution characteristics [2]. In this paper, we study the dynamic evolution of the topic sequence chain in the online social network. The paper uses the topic partition algorithm based on the mutual information and label clustering to cluster the topic labels in the social network, and get the data information in the social network. This research can help people to obtain the topic contents in each period of the dynamic evolution accurately and completely, which is very important for the further research on the dynamic social network.

The main contributions of this paper are as follows: a new method for measuring the importance of nodes is presented firstly. Based on this, a new method for calculating the class that each topic label should belong to is designed based on K-means algorithm. Experiments show that the proposed MI-LC method can significantly improve the accuracy of the topic label selection, and thus improve the accuracy and stability of the topic discovery in each period.

The rest of the paper is organized as follows: the second section introduces some related works on the topic discovery under the online social network. The third section elaborates the basic idea and the concrete steps of the proposed algorithm. The fourth section is tested by the experiments on two real datasets. The fifth section summarizes the work of this paper and looks forward to the next research direction.

2. Related Work

Hot topic detection can be traced to the task on topic detection and tracking (TDT), which is composed of the four sub-tasks that are topic segmentation, topic detection, topic tracking, and related topic discovery, among which, the topic detection task can be subdivided into online topic detection and topic review discovery [3]. Over the past ten years, a lot of social network topic discovery algorithms have been proposed out, the main points are based on the optimization of community discovery method and the heuristic-based community detection [4]. Based on the optimization method, setting the objective function and the optimal value of the iterative approximation function can realize the community discovery, and the representative method includes the spectral method and the module method [5].

In addition, there are some other effective methods. For example, in 2007, Raghavan *et al*. proposed a fast topic discovery algorithm based on label propagation [6]. The algorithm has linear time complexity and has good time efficiency in dealing with large-scale network, and does not need to optimize the number and size of the community. However, this algorithm is based on the process of iterative updating the labels of nodes, the results are often unable to achieve the expected results, and there exist uncertainty and randomness in the process of updating the node labels [7]. In recent years, different scholars have improved the standard label propagation algorithm. Literature [8] extended the label propagation algorithm and proposed a new algorithm for mining the overlapping community structure. Literature [9] proposed a label propagation algorithm based on the label influence, updating the labels of nodes when choosing the most influential label as a new node, but this algorithm does not take into account the impact of different nodes on the importance of community. Literature [10] proposed a node importance measurement method, but did not take into account the impact of neighboring node aggregation coefficient on the importance of nodes. Literature [11] proposes a node importance measurement method based on degree and clustering coefficient, but has no effect on the importance of nodes. However, to our knowledge, our work is the first attempt that employs mutual information and label clustering to discover the topics dynamically under the online social network.

3. Theoretical Foundation

3.1 Mutual Information and Self Information

In 1948, American mathematician Claude Elwood Shannon published a mathematical communication theory, which marked the production of information theory [12]. Entropy and mutual information were proposed out in this book, and these two concepts have greatly promoted the algorithm of data mining. Mutual information is a kind of entropy application form, which can describe the information between two probabilities. And mutual information can also be regarded as a random variable that contains information about another random variable, so that the dependence between the two variables can be determined by mutual information

Self-information of event $x = a_i$ in event set X is [12]:

$$I_x(a_i) = -\log P_x(a_i) \tag{1}$$

That is:

$$I(X) = -\log P(X) \tag{2}$$

Where, $a_i \in A = \{a_1, a_2, \cdots, a_n\}$, and $\sum_{i=1}^{n} P_x(a_i) = 1$, $0 \le P(X) \le 1$. Self information quantity *I* is required to be non negative value and is a random variable, $I(X)$ is a monotonically decreasing function of $P(X)$.

3.2 Information Entropy and Conditional Entropy

In order to reflect the uncertainty of the information source, information entropy is defined. Information entropy is a quantity that is characterized by the mean value of the information, which means that the average amount of information provided by each source symbol in the source output. The entropy of discrete random variable X is defined as the average value of the self information, which is represented by $H(X)$ [12].

$$\begin{aligned} H(X) &= \underset{p(x)}{E}[I(x)] = \sum_x p(x) I(x) \\ &= -\sum_x p(x) \log p(x) \end{aligned} \tag{3}$$

That is:

$$H(X) = H(p_1, p_2, \cdots p_n) \tag{4}$$

Where, $x \in X = \{x_1, x_2, \cdots, x_n\}$, $I(x)$ is the self-information of the event *x*, $\underset{p(x)}{E}$ represents the average computing of the random variable $p(x)$, $\sum_{i=1}^{n} P_i = 1$, $0 \le P_i \le 1$.

If the random variables *X* and *Y* are not independent of each other, the destination receives information *Y*, then *H(X|Y)* is used to measure the conditional entropy. After the information sink receives the random variable *Y*, the random variable *X* is still uncertain [12].

$$\begin{aligned} H(X \mid Y) &= -\sum_x \sum_y p(xy) \log p(xy) \\ &= \sum_y p(y)[-\sum_x p(x \mid y) \log p(x \mid y)] \\ &= \sum_y p(y) H(X \mid y) \end{aligned} \tag{5}$$

Where, $H(X \mid y) = -\sum_x p(x \mid y) \log p(x \mid y)$ is the entropy of *X* when *y* takes a specific value.

3.3 Average Mutual Information

Firstly, mutual information between set *Y* and event $x \in X$ is defined as follows [2]:

$$I(x;Y) = \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{p(y)} \tag{6}$$

Where, $I(\mathrm{x};\mathrm{Y})$ represents an average mutual information that is provided by event *x* about the collection *Y*.

Secondly, the average mutual information between set *X* and set *Y* is defined as [12]:

$$\begin{aligned} I(X;Y) &= \sum_{x} p(x) I(x;Y) \\ &= \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{p(y)} \\ &= \sum_{xy} p(x) p(y \mid x) \log \frac{p(y \mid x)}{\sum_{x} p(x) p(y \mid x)} \end{aligned} \tag{7}$$

3.4 Relationship between Average Mutual Information and Entropy

It is easy to prove that there exist the following relations between the average mutual information and information entropy:

$$I(X;Y) = H(X) - H(X \mid Y) \tag{8}$$

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{9}$$

$$I(X;Y) = H(X) + H(Y) - H(XY) \tag{10}$$

Relationship between average mutual information and entropy is shown in Figure1:

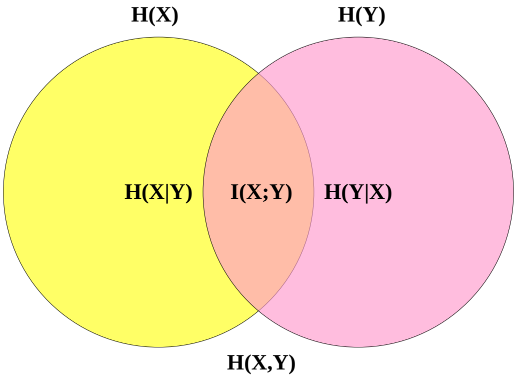


Figure 1. Relationship between Average Mutual Information and Entropy

3.5 Constructing the Topic Time Series Relation Chain

Under the online social network, when studying the topic discovery, the topic time series relation chain is firstly proposed. Suppose the topic time series chain is an undirected graph $G = (V, R_{sem}, R_{time})$, in which, the collection V contains all the nodes data under the online social network, R_{sem} represents the semantic relationship between the nodes, R_{time} denotes the temporal relations of the edges, which can be regarded as a time stamp symbol on the edge. Time series relationship chain is shown in Figure 2:

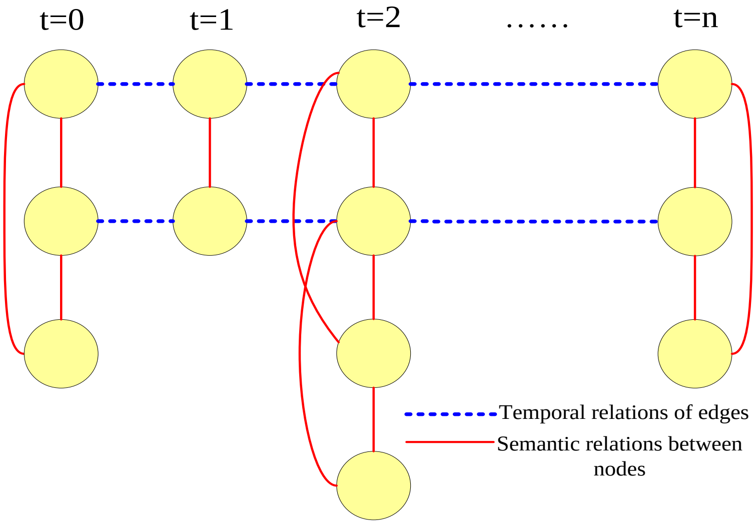


Figure 2. Graph on the Topic Time Series Chain under the Social Network

4. The Proposed Algorithm

In this paper, a new method based on mutual information and label clustering is proposed that is abbreviated as MI-LC, which is based on the measuring the node importance and clustering the labels. The relative concepts are introduced as follows:

4.1 Measuring the Node Importance

Suppose each vertex is composed of many labels and the label set is denoted as $L=\{L_1,L_2,\cdots,L_n\}$, in which, every sub-label is discrete and these discrete values can be mapped to a continuous value space $V=\{v_1,v_2,\cdots,v_m\}$, which is stipulated that $v_1<v_2<\cdots v_m$. And $V_{Ai}=\{v_{i1},v_{i2},\cdots,v_{ik}\}$ is regarded as a sub-set of $V=\{v_1,v_2,\cdots,v_m\}$, that is $V_{Ai}\subset V$.

According to the thoughts of mutual information, $V_{Ai}=\{v_{i1},v_{i2},\cdots,v_{ik}\}$ are calculated to belong to the classification $C=\{c_1,c_2,\cdots,c_t\}$, and then fill in the corresponding column L_i and row v_{ij}. The computing formula is as follows:

$$N_{imp}(L_i,v_{ij})=I(v_{ij};C)=\sum_{c_k}p(c_k\mid v_{ij})\log\frac{p(c_k\mid v_{ij})}{p(c_k)} \tag{11}$$

Where, $p(c_k\mid v_{ij})$ is the conditional probability that the value v_{ij} of attribute A_i belongs to the category c_k. Firstly, the number of samples $Num[A_i][v_{ij}]$ in the training set is computed, then $Num[A_i][v_{ij}][c_k]$ that the number of samples $Num[A_i][v_{ij}]$ belongs to the class $\mathrm{c_k}$ is also calculated, finally we obtained $p(c_k\mid v_{ij})$ as shown in the formula (12).

$$p(c_k\mid v_{ij})=\frac{Num[A_i][v_{ij}][c_k]}{Num[A_i][v_{ij}]} \tag{12}$$

4.2 Label Clustering of Vertices Based on K-Means Algorithm

The bigger the node importance, the greater the influence on other nodes, the more easily the labels of the nodes are transmitted. Therefore, when clustering the labels, the importance of nodes and the close degree of nodes should both be considered. In this paper, K-means algorithm is adopted to perform label clustering. K-means is kind of clustering algorithm based on the distance, which adopts the distance as the evaluation index, that is, the closer the distance between the two topic label, the greater the degree of similarity. The proposed algorithm on label clustering is as follows:

Input: the label clustering number *K*, as well as a matrix containing $i \times j$ topic labels.

Output: *K* topic label clustering to meet the minimum standard of variance.

The algorithm process is as follows:

(1) Randomly select the *K* topic labels from the V_{ij} topic labels as the centroids.

(2) According to the sample designation, the *K* representative sample labels are selected as the centroids of the initial class. Measuring the distance between the remaining topic labels with each centroid, and dividing the topic labels into the class including the nearest centroid point. The adopted distance calculation method is Euclidean distance. Euclidean distance is computed as follows:

$$V = \sum_{i=1}^{k} \sum_{x_j \in V_j} (\mathrm{x}_j - \mu_i)^2 \tag{13}$$

Where, μ_i denotes the *ith* clustering center, x_j is every topic label in the vertex V_j.

(3) Recompute the centroid of each class which has been obtained. Redetermine the *K* class centroids and calculate the mean value of the *K* variables, and the mean points are regarded as the centroids of the *K* class.

(4) Step 2 and step 3 are iterativly performed until the new centroid is equal to the original centroid or less than a given threshold, the algorithm terminates.

4.3 Algorithm Implementation

Combined the nodes importance with the label similarity, the proposed topic discovery algorithm MI-LC based on mutual information and label clustering is designed and the detailed algorithm is as follows:

(1) Initialize the topic labels of each vertex $v \in V$;

(2) According to the formula (11), the importance of each vertex is calculated, and the nodes are sorted from high to low according to the importance of nodes and produce the ordered lists $V' = \{v_1, v_s, \cdots, v_n\}$, among them, $S(v_1) \geq S(v_s) \geq \cdots S(v_n)$;

(3) Iteration number *t=1*;

(4) For any $v_i \in V'$, the topic label is updated as its nearest neighbor label with k-means algorithm based on Section 4.2:

According to the formula (13), the iteration and updating process of topic labels are executed. During the process of iteration, each vertex adopts the most adjacent centroid to update its own label. If there exist two or more centroids for the equal nearest value, then randomly selecting one as the label of the vertex.

The topic labels are updated asynchronously, each topic label is jointly determined by the iteration in time *t* and the iteration in time *t-1*, the formula is as follows:

$$\begin{aligned} &L_x(t) = f(L_{xm1}(t-1), \cdots, L_{xmn}(t-1)), \\ &L_{xm(n+1)}(t), \cdots, L_{xmk}(t)), x_m \in ver(x) \end{aligned} \tag{17}$$

Where, $L_x(t)$ is the topic label of the vertex x in the iteration in time *t*, $ver(x)$ is the adjacent vertex set of vertex x.

(5) If the number of iterations $t == \max Iter$ or the topic label of each vertex is the maximum value of the label, the vertex with the same label is placed in the same topic, and the algorithm terminates; otherwise $t = t+1$, then return step (4).

5. Experimental Results and Analysis

5.1 Experimental Datasets and Experimental Environment

In order to verify the effectiveness of the proposed MI-LC algorithm, we select two different real social networks, which are DBLP and People's news network (http://www.peoplenews.com.cn/), respectively. The first data set extracted from DBLP is a paper collaboration network on computer science field, each vertex represents an author of the paper and there exist collaborative relationship between the authors. The second data from People's news network is a news sharing social network, in which, each vertex denotes the corresponding user and each edge represents the relationship between the vertices. The detailed information on two data sets is shown in Table 1:

Table 1. The Detailed Information of Two Experimental Datasets

Dataset	Vertex	Edge	Type
DBLP	262,056	1,023,241	Collaborative net
People News Net	183,523	981,561	Social Network

Inter (R) Core (TM), CPU@3.30Ghz, i3-2120 and RAM 3GB are configured as hardware environment, and the software adopts Windows Microsoft 7.0. All algorithms are implemented with Java language and are tested on the JDK 7 platform.

5.2 Evaluation Metrics

5.2.1 Normalized Mutual Information NMI

NMI (Normalized Mutual Information) is a method based on the information theory to measure the similarity of two methods, which is widely used and has the reliability [13]. In the experiment, LPA [6] algorithm and LDA [14] algorithm are compared in NMI value and running time. In order to overcome the randomness of the algorithm, the average value of the results of all the experiments are taken for the 50 times. The maximum iteration number of the algorithm is 100. NMI is defined as:

$$NMI(X|Y) = 1 - [H(X|Y) + H(Y|X)]/2 \tag{9}$$

Among them, *X* is a collection of all real topics under online social network, Y is a collection of all prediction topics, $H(X|Y)$ is the normative conditional entropy of X on Y.

5.2.2 Perplexity Value

The model can be well modeled in the existing data, but whether it has the same effect on the non observational data, which depends on their generalization ability. The perplexity degree of a discrete probability distribution is defined as [15]:

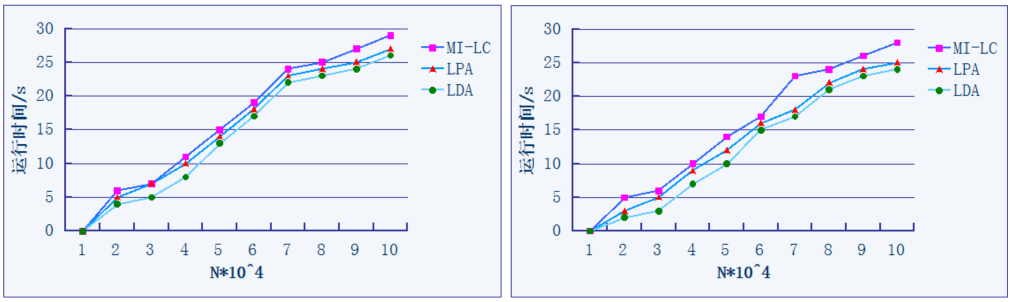
$$2^{H(p)} = 2^{-\sum_x p(x)\log_2^{p(x)}} \tag{10}$$

Where, *H(p)* is the entropy of the distribution of the events in *x* ranges. Perplexity of a random variable *x* may be defined as the perplexity of the distribution over its possible values *x*.

5.3 Experimental Results Analysis

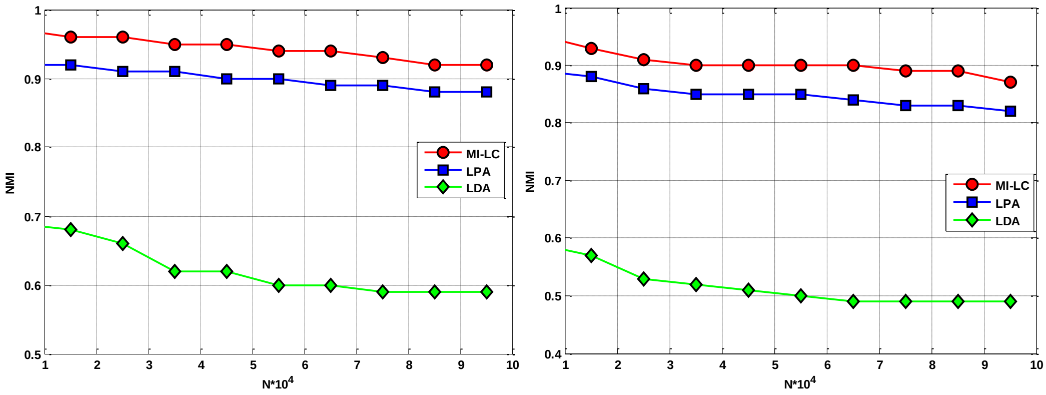
5.3.1 Experimental Results Analysis on NMI Value

The operating time and NMI value of different algorithms are shown in Figure3 and Figure4 with the change of the amount of nodes. From Figure3, it can be observed that, similar with LPA and LDA, there also exists an approximate linear relationship between running time and the number of nodes in the proposed MI-LC algorithm under datasets DBLP and people's news network. The importance of nodes needs to be calculated and the label clustering needs to be processed, which result in that the time efficiency of the algorithm is slightly lower than the LPA and LDA. Although the time efficiency of LPA and LDA algorithm is slightly better than the MI-LC algorithm, it can be seen from Figure4 that they are inferior to the MI-LC algorithm in the quality and stability on topic discovery. This is mainly due to during the process of topic discovery, the close degree between nodes and the importance of nodes are comprehensively considered, which improve the stability and accuracy of the algorithm compared with LPA and LDA.



(a) On DBLP Dataset (b) On People's News Network Dataset

Figure3. Operating Time Comparison of MI_LC, LPA and LDA



(a) On DBLP Dataset (b) On People's News Network Dataset

Figure 4. Comparison of NMI Value on MI_LC, LPA and LDA

5.3.2 Model Generalization Capability Analysis about Perplexity Value

Here we randomly divided the original data into five parts, each of which takes one of part as the test set, and the rest as the training set. According to the number of topics, the average perplexity values are calculated in five cases. Table 2 and Table 3 provide the perplexity values in the number of *K* corresponding to the topic. From Table 2 and Table 3, it can be seen that the perplexity values of MI-LC algorithm outperform those of LPA and LDA, which shows that the generalization ability of MI-LC model is better than that of LPA and LDA. Otherwise, the perplexity values of LPA are also slightly better than LDA model. In comparison, the topic discovery model based MI-LC algorithm has a better generalization ability for non observational data compared with LPA and LDA model.

Table 2. Perplexity Value Comparison among MI-LC, LPA and LDA on DBLP Dataset

Models K	MI-LC	LPA	LDA
10	2534.21	2346.18	2305.03
20	2413.78	2215.56	2145.89
30	2267.23	2136.32	2005.29
40	2109.16	2001.23	1999.34
50	1999.34	1997.26	1996.06

Table 3. Perplexity Value Comparison among MI-LC, LPA and LDA on People's News Dataset

Models K	MI-LC	LPA	LDA
10	2682.15	2476.11	2214.10
20	2588.37	2381.52	2008.56
30	2459.01	2275.78	1999.05
40	2223.56	2193.17	1997.61
50	2135.68	2007.35	1994.95

In a word, on the one hand, MI_LC algorithm considers the importance of nodes, which makes the nodes with higher importance impact the nodes with lower importance. On the other hand, MI_LC algorithm takes into account the label clustering, which improves the accuracy of label selection, the quality and stability of the topic discovery under the social network.

6. Conclusions and Future Work

In this paper, we investigated the problems of the traditional label propagation algorithm and proposed a new reliable topic discovery algorithm based on mutual information and label clustering named MI-LC under dynamic social network. Through theoretical analysis and comparative experiments on two real data sets, the proposed algorithm MI-LC can significantly improve the accuracy and stability of topic discovery. However, the proposed MI-LC algorithm has some shortcomings that need to be improved, which is also our next research direction. The existed disadvantages are as follows: Firstly, during the process of label clustering, the number of initial clustering centers *K* should be given in advance. It is very difficult to estimate *K* value, and how to select the appropriate *K* is one of our research directions. Secondly, from the framework of MI-LC algorithm, it can be seen that the algorithm needs to be adjusted the sample classification, so when the volume of data is very large and the time cost is very large, the time complexity of the algorithm needs to be analyzed and improved.

Acknowledgements

This work was supported by the Scientific Research Innovation Program for College Graduates of Jiangsu Province in China (NO. KYLX15_0321), the Key Project of Overseas Visiting for the Outstanding Young Backbone Talents at Universities and Colleges of Anhui Province in China (NO. gxfxZD2016267) , the Key Project of Anhui Province Colleges and Universities Natural Science Foundation of China (No.KJ2014A250) and the Project of Outstanding Young Talents supported by Suzhou University of China (No.2014XQNRL009, No.2014XQNRL006, No.2014XQNRL007).

References

[1] S. Saganowski, P. Bródka and P. Kazienko, "Influence of the dynamic social network timeframe type and size on the group evolution discovery", IEEE/ACM Inter. Conf. on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Istanbul, Turkey, **(2012)**.

[2] B. Hu, Z. Song and M. Ester, "User features and social networks for topic modeling in online social media", IEEE/ACM Inter. Conf. on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Istanbul, Turkey, **(2012)**.

[3] X. Wei, J. Sun and X. Wang, "Dynamic mixture models for multiple time series", Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, **(2007)**.

[4] R. M. Nallapati, W. Cohen and S. Ditmore, "Multi-scale topic tomography", Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, **(2007)**.

[5] X. R. Wang and A. McCallum, "Topic over time: a non-markov continuous-time model of topical trends", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, **(2006)**.

[6] U. N. Raghavan, R. Albert and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical review. E, Statistical, nonlinear, and soft matter physic, vol. 3, **(2007)**.

[7] J. Xie and K. S. Boleslaw, "Community detection using a neighborhood strength driven label propagation algorithm", Proceedings of the 1st IEEE Network Science Workshop, IEEE Computer Society, Washington, DC, USA, **(2011)**.

[8] G. Zhang, J. Zhang, J. Yang, Y. Xin and S. Wang, "An Overlapping Community Structure Detecting Algorithm Based on Gaussian Field Label Propagation", Journal of Computational Information Systems, vol. 10, **(2014)**, pp. 4331-4338.

[9] H. Shi, Y. Song, L. Shi and C. Shu, "A high-influence greedy maximization algorithm based on community structure", Journal of Computational Information Systems, vol. 2, **(2015)**, pp. 449-456.

[10] M. He, M. Leng, F. Li, Y. Yao and X. Chen, "A node importance based label propagation approach for community detection", Knowledge Engineering and Management, vol. 214, **(2014)**, pp. 249-257.

[11] Z. Ren, F. Shao, J. Liu, Q. Guo and B. Wang, "Node importance measurement based on the degree and clustering coefficient information", Acta Physica Sinica, vol. 12, **(2013)**, pp. 505-515.

[12] J. Lin, "Divergence measures based on the Shannon Entropy", IEEE Transactions on Information Theory, vol. 1, **(1991)**, pp. 145-151.

[13] P. A. Estévez, M. Tesmer, C. A. Perez and J. M. Zurada, "Normalized mutual information feature selection", IEEE Transactions on Neural Networks, vol. 2, **(2009)**, pp. 189-201.

[14] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993-1022.

[15] K. Salomatin, Y. Yang and A. Lad, "Multi-field correlated topic modeling", Proceedings of the SIAM International Conference on Data Mining, Sparks, Nevada, USA, **(2009)**.