Advances in Social Science, Education and Humanities Research (ASSEHR), volume 75
2016 International Seminar on Education, Innovation and Economic Management (SEIEM 2016)

Evolution Analysis of Topics on Social Media Based on the Co-word Network

Chen Zhuoqun
School of Media Science
Northeast Normal University,
Jingyue Street No.2555
Changchun, Jilin Province, China
hilda1006@163.com

Sun Xu
School of Economics and Management
Jilin Jianzhu University
Xincheng Street No. 5088,
Changchun, Jilin Province, China

***Abstract*—For the information in the social media, methods of topic feature selection in different time have been put forward to build the dynamic co-word network. The community discovery algorithm is applied to divide the co-word network on the basis of communities. And the co-word network community stands for the subtopics of the topics in the social media. On the basis of the comparability identification of subtopics in different time, the evolution process of subtopics is divided into three stages as the subtopic production, the subtopic diffusion and the subtopic fading. The empirical analysis shows that representing subtopics by using the co-word network community has such advantages as the intelligibility and the noise reduction; the development path and variation trend of subtopics can be clearly clarified on the basis of the co-word network community.**

Keywords—Co-word Network; Community; Social Media; Topic Identification; Topic Evolution

I. INTRODUCTION

With the further development of Web2.0, a large number of interactive online communities focused on users have emerged on the Internet, whose features have both sociability and transmissibility of media. Typical applications include SNS websites (Social Networking, like Renren and Qzone), microblogs, Wechat, forums and others[1]. Currently, the academia names those network applications in many ways, like social media, socialized media and social mention, and the concept of social media is utilized in this thesis. In the social media, an enormous amount of User Generated Content (UGC) is produced in forms of writing, sharing, making comments, discussion and others by users, among which includes not only the information related to the daily life of users, but also a large number of opinions on social events and enterprise brands given by users. The social media has become an important place where the online public sentiment is produced, diffused and discussed.

The social media is important data sources for the analysis of the public sentiment. Researching the social media can comparatively and truly get the subjective information from online users. The social media contains comparatively abundant contextual metadata, like when and where the user posts the information. Researching the contextual metadata can also reflect the contextual information on the topic of public sentiment in some way. In the discussion in the social media, concerns of a topic can be changed with time. In combination with when and what the user post, the development trend of the public sentiment can be known by researching and analyzing different concerns of one same topic in different time in the social media, which can be used for the evolution analysis of the public sentiment, like the topic about *offsite college entrance examination* in March 2013. The information reported first was that *children of postdoctors in Beijing can take offsite college entrance examination*. And then some derivative subtopics are reported, such as *regional inequality in offsite college entrance examination* (Taking the college entrance examination in another place instead of in the hometown is not fair because the policies for the college entrance examination are different in different places in China.), n*ational college entrance exam migrant* (In order to enter into an university or a better university, people move to the provinces whose entrance marks are lower and entrance rates are higher.) and *admission ticket bought in offsite* (In order to have the qualification of taking the college entrance examination in the places that enjoy better policies for the examination, people bought houses there.)[2]. The main goal of the evolution analysis of topics is to analyze the construction of subtopics, the importance of subtopics, the developmental trend of subtopics, topic migration and others in different time[3]. In the field of analysis on scientific and technological information, the co-keyword network is often used to represent the topics in the field of science and technology[3-5], while in the field of text mining, the textual topic is often represented with the aid of the vector space model[6, 7], the topic model[8] and others of subject document collection. In this thesis, the co-keyword network is generalized to the general co-word network in combination with the textual features of the social media and on the basis of that, subtopics are expressed by using the co-word network community to explore and analyze the evolution of subtopics, realizing the evolution analysis of topics of social media.

II. RECOGNITION OF SUBTOPICS BASED ON CO-WORD NETWORK COMMUNITY

The topic evolution analysis is an important research direction in the research of topic surveillance and tracking, including the recognition of subtopics, the judgment of the relation among subtopics, the attention assessment of subtopics, the analysis of evolution and other subtasks[6]. The

This paper is the medium-term outcomes for Ministry of Education Humanities and Social Sciences Research Youth Fund Project "Research on Knowledge Transfer and Dissemination Based on Academic Blog "

recognition of subtopics means that topics in topic document collection are further divided into subtopics, and the subdocument collection (subtopic document collection) related to subtopics are found out in document collection, representing subtopics in some forms (For instance, subtopics are represented by co-keywords). In this section, for different time of topics, feature words of topics are recognized to establish the dynamic co-word network in different period of time, and subtopics are found out and represented on the basis of that.

A. Recognition of Topic Feature Words

Massive information in the social media is written by common users, whose words are more colloquial and are also with informal forms of emoticons, cyberword, etc. A topic feature word in the social media refers to the word which can reflect the topic content and has definite meaning in the content made by users. In the view of the types of vocabularies, in this thesis, topic feature words are restricted to such polysemous words as nouns, verbs and adjectives, removing the words as adverbs, pronouns, onomatopoetic words and various cyberword and symbols.

As time goes on, the attention of the public sentiment (namely, the subtopic) is often transferred, and from the aspect of feature words, topic feature words in different periods of time will change. In the process of recognizing topic feature words, feature words should be pointedly selected from the topic document collection in different periods of time. In this thesis, the significant priority of feature words is calculated by using TFIDF method[9] and on the basis of that, topic feature words in different period of time are selected. The weight calculation formula of the feature word *w* is as follows:

$$\mathrm{tfidf}_{\mathrm{ts}}(\mathrm{w}) = \sum_{\mathrm{deD_ts}} (\mathrm{tf}_{\mathrm{w,d}} + 0.5) * \log \frac{\mathrm{N} + 0.5}{\mathrm{df}_{\mathrm{w,D_{ts}}} + 0.5}$$

Among which, D_{ts} refers to the document collection in the topic period of time (ts) ; $\mathrm{tf}_{\mathrm{w,d}}$ refers to the occurrence frequency of *w* in document *d*; $\mathrm{df}_{\mathrm{w,D_{ts}}}$ refers to the number of documents in which *w* appears; *N* refers to the number of documents for the whole topic document collection. Through the weight value, the words which appear frequently in different periods of time while appear less frequently in the whole document collection can be selected as the feature words for topics in one period of time.

B. Construction of the Dynamic Co-word Network

In the field of scientific and technological information, the co-word network is believed to present the cognitive structure of the subject[3]. The cognitive meaning of the co-word network can also be extended to the broader field of textual mining. For instance, when the analysis of the public sentiment is applied, the co-word network which is based on common words is an effective method to find out and visually present the topics of the public sentiment[10]. Similarly, the co-word network is used to dig the topics of the public sentiment and to analyze its topic evolution process in this thesis.

According to the co-occurrence relation of feature words in the same text window, the co-word network is established after the topic feature words in different periods of time being recognized. The granularity of the text window can be the levels of sentences, paragraphs, files and others. The corresponding topic co-word network is a dynamic co-word network and it can be cut into slices in accordance with the periods of time of subtopics because the data of the topic document collection has time attributes. DG, the dynamic co-word network of the public sentiment is represented as: $\left\{ <\mathrm{G}_1, \mathrm{ts}_1>, <\mathrm{G}_2, \mathrm{ts}_2>, \ldots, <\mathrm{G}_\mathrm{i}, \mathrm{ts}_\mathrm{i}>, \ldots \right\}$, among which, ts_i refers to topic period of time (i), G_i refers to the co-word network in i-topic period of time. The nodes of the topic co-word network represent topic feature words and the edge represents the co-occurrence relation of feature words in the same text window. Through establishing co-word networks for different periods of time, semantic features of vocabularies are remained and the internal relationship among vocabularies is carried out, which can express richer information. In fact, the co-word network of the public sentiment topics is a new representation model of texts--representation model of text map[11]. In the representation model of text map, the feature words of text are not independent with each other, but connected by edges, which represents the internal relationship between words, so the information carried is richer. Potential models can be found out by digging and analyzing the structure of the dynamic co-word network with the network mining algorithm.

C. Recognition of Subtopics Discovered on the Basis of Communities

According to language habits, one same topic or theme is often expressed by some core vocabularies, which is widely applied in the Topic Model[12] in the current machine learning field. Based on the feature of topic phrases, the enlightenment can be gotten as follows: Through digging the structure of the co-word network, can the core vocabulary set which is used to express some topic of the public sentiment be found out or not?

In the fields of the complex network, the social network and other networks, Community is an important structural feature between the macro and micro levels[13]. In many realistic networks, networks are not even or completely random, which often shows the assemblage of parts of nodes because of some reason. The connection between the nodes is comparatively close so that the assemblage of nodes is taken on the network topology. In the social network, this phenomenon is presented as groups, communities and others composed of people; in the citation network, this phenomenon may be presented as a collection of research papers connected by citation relations of papers in some research field; in the biological network, this phenomenon may be presented as biological units composed of concrete similar functions. In these groups, there are some connections among nodes in communities so that they can be closely united and present partial common features. If the co-word network is considered as a complex network structure, and in the co-word network, the community structure closely connected by those nodes can be found out, the vocabularies in communities can be understood as a collection of vocabularies belonging to the same cognitive structure[10]. Therefore, communities in the co-word network reflect the themes in their corresponding document collections. Accordingly, in the document collections of the public sentiment topics, these communities are the subtopics of the topics of the public sentiment and belong to the different attention of the topics of the public sentiment.

At present, there are many mainstream community discovery algorithms, like the Kernighan-Lin algorithm, spectrum dichotomy, the hierarchical clustering method[14], but the lacks of those algorithms are that they only apply to deal with the small networks which have less nodes while they have higher time complexity and lower applicability for massive networks which have more nodes. The Louvain algorithm[15] is adopted to discover the community structure in the co-word network in this thesis. The Louvain algorithm is a heuristic algorithm with optimization based on modularity, whose characteristics are as follows: (1) communities are hierarchically divided from the bottom to the top, which is in accordance with objective laws, and the results have stronger explanations; (2) the number of nodes supported by calculation is larger, which can apply to the division of communities of massive networks, and the efficiency is higher. In the process of implementation, for the co-word network in different topic time, different subtopics discussed in the corresponding period of time are recognized by the application of the community discovery algorithm.

III. Exploration of Topic Evolution Based on Co-word Network Communities

In the social media, there may be same subtopics among topics in different periods of time. The form is presented as different communities in different co-word networks. Similar subtopics are recognized by adopting some method and then they are conducted by the evolution analysis. Through observing different community structures in different co-word networks, the similarity among subtopics in different time can be judged. Based on that, evolution types of subtopics are obtained and the method of conducting evolution exploration by using co-word network communities is also obtained.

A. Recognition of the Similarity of Subtopics in Different Periods of Time

The information in different time in the social media may have some subtopics which are discussed constantly and may also have more similar subtopics. In the process of analyzing the evolution of topics, first, the association relationship among the subtopics in different time should be recognized to explore the development paths of subtopics. In the methods of text mining, the similarity judgment methods of topics can be summarized as the Jaccard method and the cosine similarity that are based on the bag-of-words model, the KL distance method and the theme model method that are based on the language model, and others. Different from these methods, it is a method for text map to represent subtopics by using the co-word network community.

According to the features in representation model of text map, the similarity of subtopics turns to the similarity of network structure in co-word network community. Structurally, text map represents two kinds of structural elements, namely points and edges, in models. Accordingly, the similarity between two text maps is often calculated through the matching degree of nodes and edges[17]. Specifically, the recognition of similarity can be conducted by using such matching algorithms based on the structures of figures as the method of similar size, similar capacity method, edge weights method and the largest public subgraph method[11].

After the recognition of the co-network communities in different time, the similarity between two different communities in different time is calculated by using the similarity degree calculating methods based on text map representation model and a concrete threshold is set up to make sure whether two different communities are similar or not. In that case, the similarity between subtopics is judged and same or relevant subtopics in different time are recognized.

B. Evolution Process of Subtopics Based on Co-word Network Communities

According to the rules of the evolution of the complex network community, the evolution of co-word network communities can be presented as such types of evolution as the production, the fading, the division, the merger, the expansion and the contraction[16]. In combination with the characteristics of topics in the social media, the evolution of subtopics life cycle can be divided into three stages, namely, the subtopic production, the subtopic diffusion and the subtopic fading. The definition and the method of recognition in each stage are as follows:

1) Subtopic production: this stage is the initial stage for the subtopics of the public sentiment in the social media. It may occur in the initial stage of some topic in the social media and it may also occur in the initial stage of the subtopics which are derived after the production of some topic. As the subtopics which are produced in the initial stage, namely produced in the period of time of the first topic, they can be presented in accordance with the communities of the co-word networks in the time of the first topic. In the current time, the subtopics which are derived from topics may not exist in the co-word network in the former period of time and may only belong to a part of other subtopics, while they have already become an individual co-word network community in the co-word networks in the current period of time. The part of the subtopics in the current period of time, which belongs to the subtopics in the former period of time, is judged from whether the community nodes of the current subtopics are only a part of the communities of the subtopics in the former period of time.

2) Subtopic diffusion: in this stage, with the rise of the attention of subtopics, the number of relevant documents issued by users in the social media is increasing. At the same time, the feature words of subtopics are increasing, too, which is presented as the increase of the data of nodes and the strength of the edge weight in the co-word network communities that are corresponding to subtopics. The method of recognition is to compare the number of nodes and edge weights in the co-word network community in the current period of time with those in the corresponding co-word network community in the former period of time. The greater degree of diffusion of subtopics, the more increase of the number of nodes and the more strength of the edge weight.

3) Subtopic fading: in this stage, with the decrease of the attention of subtopics, the number of documents issued by users is declining. At the same time, the feature words of subtopics are decreasing, too, which is presented as the decrease of nodes and strength of the edge weight in the co-word network communities that are corresponding to subtopics. The reason why subtopics are fading may be that the attention of subtopics in the public is decreasing or the topic is drifted, namely, the public have distracted their attention from one subtopic to another. The method of recognizing the fade of subtopics, which is similar with the method of recognizing the diffusion of subtopics, is to compare the changes of the number of nodes and the edge weight in the co-word network communities between in the current period of time and in the former period of time. If the number of nodes and the strength of the edge weight decrease faster, subtopics will fade faster.

IV. CONCLUSION

In this thesis, through analyzing the social media by the co-word network method, the extract method of feature words in different periods of time and the method of building dynamic co-word network are proposed, and the social division is taken to the co-word network by the community discovery algorithm. The subtopic of the social media topic is represented by co-word network communities. Based on the similarity identification of subtopic in different periods of time, the subtopic evolution is divided into three stages as the subtopic generating, the subtopic diffusion and the subtopic fading. Based on the co-word network communities, the development paths and variation tendency can be clarified clearly, and the core subtopic in different periods of time can also be analyzed, so as to take the evolution analysis on the topic of social media. There are still inadequacies in the methods in this thesis: the time division of topic can not be automatically identified according to the feature of document sets; the influences from extract methods of different features to topic evolution analysis are not compared; the further study for the reliability and further refining methods of subtopic are required.

REFERENCES

[1] Zhang Qiugui. Shaping and Spread of Publishing Brand under Social Media Environment[J]. Publishing Research, 2013(01): 12-14.

[2] Huang Weidong, Chen Lingyun, Wu Meirong. Emotional Evolution Research on Topics of the Public Sentiment[J]. Journal of Intelligence, 2014(01): 102-107.

[3] Wang Xiaoguang. Formation and Evolution of Scientific Knowledge Network (I): The Method Proposing of Co-word Network[J]. Journal of Information, 2009, 28(4): 599-605.

[4] Liu Zeyuan, Yin Lichun. The Visualization Research on Co-word Network of International Science Subject [J]. Journal of Information, 2006,25(5): 634-640.

[5] Feng Lu, Leng Fuhai. Theoretical Progress of Co-word Analysis Method[J]. China Library Journal, 2006, 32(2): 88-92.

[6] Wang Wei. The Evolution Analysis of Network Topic Based on Key words and Point-in-time[D]. Shanghai: Fudan University, 2009.

[7] Long Zhiwei, Cheng Wei. Detecting Algorithm of Hot Topic Based on Word Clustering[J]. Computer Engineering and Design, 2011(06): 2214-2217.

[8] Shan Bin, Li Fang. Method Review Based on the Evolution of LDA Topic[J]. Journal of Chinese Information Processing, 2010(06): 43-49.

[9] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press Cambridge, 2008.

[10] Tang Xiaobo, Song Chengwei. The Analysis of the Microblog Public Sentiment Based on Complex Network[J]. Journal of Information, 2012, 31(11): 1153-1162.

[11] Li Gang, Mao Jin. Representation Model of Text Map and Its Use in Text Mining[J]. Journal of Information, 2013, 32(12): 1257-1264.

[12] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[13] Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks[J]. Physical Review E, 2004, 69(2): 26113.

[14] Li Gang, Ren Jiajia, Mao Jin, Yang Guancan, Community Structure Analysis of The Patentee Cooperation Network[J]. Journal of Information, 2014, 33(3): 267-276.

[15] Blondel V D, Guillaume J, Lambiotte R, et al. Fast Unfolding of Communities in Large Networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008.

[16] Wang Xiaoguang, Cheng Qikai. Visual Analysis of Subject Theme Evolution Based on NEViewer[J]. Journal of Information, 2013, 32(9): 900-911.