A Community Discovering Method Based on Event Network for Topic Detection

Tong WAN*, Wei LIU*, Zongtian LIU*

*School of Computer Engineering and Science, Shanghai University, China

lavan@live.cn, liuw@shu.edu.cn

Abstract— **Traditional text modeling methods are mainly based on word frequency statistics, which lacks of necessary semantic information, and in real-time news topic tracking, it is difficult to update topics to catch up with their growth and variation. In this paper, we use event network to model news text. Through using community discovering algorithm in event network, we can obtain event cluster, and accordingly achieve topic detection. Event network is a weighted directed network, therefore general community discovering methods can't be used directly on event network. The communities in event network are more likely to be fine granularity community, and their amount is not known in advance. Hence, we proposed a hierarchical community discovering algorithm based on event network, which exploits the semantic properties of event nodes and edge-weight information in the network, to discover fine granularity communities that are semantically meaningful. Experiment results show that the algorithm is effective. Our work is also the basis of topic merging, topic tracking and information discovering based on event network.**

Keywords— **Event Similarity, Event Network, Topic Detection, ENCDA**

I. Introduction

Internet has become the "fourth-largest media", after three traditional media: newspaper, radio and television. However, when we enjoy the advantages bring by its unique timely, massive and interactive features, we also have to endure the negative effect that it has become more and more difficult to access useful information effectively. The information presented to readers are scattered and disordered, mostly readers can't get a complete understanding of the whole situation about topics they want to know. Therefore, topic detection and tracking (TDT) received more and more attention from researchers. The main task of TDT is to discover unknown topics under the condition of absence of priori knowledge, and tracking follow-up reports of exposed topics. It can help people access information more effectively, thus have a complete understanding of details and the evolution progress of a topic[1].

TDT is mainly focused on news information recognition, data mining and organization for news articles. TDT makes it possible for computer to filter information from the Internet, thereby improving the efficiency of people to obtain the useful information. Text representation model is the basic problem of information processing technology including TDT. Traditional text modeling method, such as SVM, is mainly based on word frequency statistics, which ignores the effect of the order of text elements and the relations of them, each lexical item is assumed to be independent. However the semantics of texts not only contains in the word frequency, but also are related to the text structure and element relations. To solve this problem, in this paper we present a hierarchical community discovering algorithm based on event network. An event network is a weighted directed network that consists of a set of nodes and edges. The nodes are events, and the edges are event relations. We convert news text to event network first, by using community discovering algorithm on event network, considering both topological structure information of the network and semantic information of event nodes, we can obtain event cluster, and accordingly achieve topic detection.

This paper is organized as follows. Section 2 discusses related work of community discovering. Section 3 discusses the definitions of event, event network and community. Then, in Section 4 we present the hierarchical community discovering algorithm. Experiment results are presented in Section 5. The paper is end up with Section 6 where we present our conclusions and possible future work.

II. Related Work

In topic detection and tracking field, a news text usually contains a lot of topics, and each topic involves many events, each event has rich semantic information, including but not limited to time, object, environment, action, etc. And there are many kinds of relationships between events. Using event as basic element in news text modeling, can be able to represent the structure of text and information better than using traditional text modeling methods. Consequently, D. Wang proposed a novel semantic-based and event-oriented text representation model - event network[2], which can provide support for text semantic information processing.

General community partitioning algorithms are mostly divided into two kinds: graph theory based algorithm and sociology based algorithm. The most typical graph theory based algorithms are the Kernighan-Lin algorithm[3] and spectral bisection method. The Kernighan-Lin algorithm is an exploratory optimization method, it is essentially a greedy dichotomy algorithm based on principles to divide the network into two communities with known size. Thus, it has limitations in practical use. For spectral bisection method, either traditional spectral bisection algorithm based on Laplace matrix[4,5] or other spectral bisection method needs to know the number of communities in advance, so they can not be used for community discovering in networks whose community number is unknown. Besides there is a clique percolation method[6], of which the community partition results are influenced by a parameter K.

Sociology based community discovering algorithm is divided into division method and aggregation method. GN algorithm[7] is a division method, its fundamental principle is to get communities by finding the edge with the highest score of betweenness and remove it from the network. Although the accuracy of GN is high, but it's unable to know which step to stop if the number of community is not given. Newman proposed a fast aggregation algorithm[8], which has similar accuracy with GN, and the complexity has been significantly improved. CNM algorithm[9] is a greedy method, also proposed by Newman, which optimized the fast aggregation algorithm, and have even lower complexity. Blondel noticed that in most large networks there are several natural organization levels, communities are divided into sub-communities, so they proposed a hierarchical community detection method[10]. This method is based on modularity optimization. It has an advantage in terms of time complexity, and the quality of the communities detected is good as measured by community modularity.

Event network is a weighted directed network, each node in the network is of rich semantic attributes, each edge represent for a type of event relation. Yet those community discovering algorithms above are used in non-weighted network, and mostly based on network topological structure information, without taking the semantic properties of the node and edge weights into account. Meanwhile the communities in event network are more likely to be fine granularity community, and their amount is not known in advance. D. Wang[2] proposed a community discovering method on event network based on minimum spanning tree, which only considered event relations by calculating the appearance frequency of event pair, in spite of the rich semantic properties within each event node , and its efficiency is not high enough when applied on large network. Meanwhile, it is a top-down division algorithm, every node in network will be categorized into a community, which may be inappropriate. Furthermore, it causes another serious flaw. In TDT, topic tracking is another important task after we finished topic detection. While a news topic is continuously developing over time, more and more events happened, we need to update the community partition result to catch up with the growth and variation of topic. The aforementioned method can not solve this problem nicely, because each time this algorithm has to start all over again on the entire newly built event network, including old events and new events, and this will greatly affect the efficiency. Hence, in this paper we proposed an aggregation algorithm of community discovering on event network, which exploits the semantic properties of event nodes and edge-weight information in the network, to discover communities that are semantically meaningful from event network. Results of comparison experiments confirmed the effectiveness of our algorithm. This work is also the basis of topic merging, topic tracking and information discovering based on event network.

III. Event Network And Community

Definitions of event network presented in reference[2] is the foundation of our work. In this definition, different type of edge represents different event relation. However, edge weight are defined using statistical method, weight value is proportional to the appearance frequency of the event pair, which is unstable and not reasonable enough. In this paper, weight value of each edge is decided by the semantic similarity of the two nodes linked by the edge. In this section, we will introduce the definitions of event network, and then in section 4, we will present our discover communities algorithm.

A. Event

Definition 1 (Event): event is defined as a thing happens in a certain time and environment, which some actors take part in and show some action features. Event *e* is defined as a sextuple:

$$e ::=_{def} < A, O, T, V, \mathrm{P}, \mathrm{L} >$$

In this definition, *A* represents the action set of an event. *O* represents the objects participated in the event. *T* represents the time that an event happens, or time period during which the event lasts. *V* represents the environment in which the event happens; *P* represents the assertions on the procedure of actions execution in an event; *L* represents the language expressions. In this paper we use elements *A*, *O*, *V* and *T* to represent an event.

B. Event Network

An event network is a network that consists of a set of nodes and edges. The nodes are events, and the edges are event relations. Event instances in news text are connected together by event relations to represent the text.

Definition 2 (Event Network): an event network (EN) is a weighted directed acyclic graph that consists of a set of nodes and edges. The nodes are events, and the edges are event relations.

$$EN::=_{def}<Events,Edges>;$$

$$Events=\{e_1,e_2,\cdots,e_n\};$$

$$Edges=\{<e_i,e_j,r_{ij}>,<e_x,e_y,r_{xy}>,\cdots\}(1\le i,j,x,y\le n);$$

$$r=\{Correlation,Causal,Accompany,Follow\};$$

In this definition, e_i is an event, r_{ij} is the relation between e_i and e_j . There are four kinds of event relation type, including: Correlation, Causal, Accompany and Follow. In this paper, different relation types are treated equally in community discovering in order to simplify the process. We only consider the edge weights, which is decided by the similarity of the two event node linked by the edge.

C. Community

In event network based text representation model, text is represented by event network, and topic detection is to partition several closely interrelated set of events from the network. Here we borrow the concept of community in social network. Network community is virtually a description of the closeness between nodes. Nodes in the same community are closely interrelated, while nodes between different communities are distant.

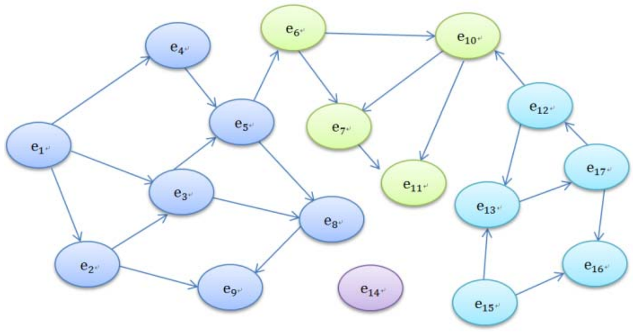

Figure 1. Communities in EventNetwork

The most commonly used community partition quality measure on non-weighted network is modularity. Define graph $G = \{V,E\}$, G has n nodes and m edges. On non-weighted undirected graph, modularity Q[11] is defined as:

$$Q_{non-weighted} = \frac{1}{2M}\sum_{i,j}^{n}(a_{ij} - \frac{k_i k_j}{2M})\delta(C_i, C_j)$$

In this definition, k_i is the degree of V_i , C_i is the community that node i belongs to, $\delta(C_i,C_j) = 1$ when node i and node j belong to the same community， otherwise $\delta(C_i,C_j) = 0$. The main function of modularity is to evaluate the community partition quality and the structure strength of network communities. Q value ranges from 0 to 1, higher value means better community structure strength. Generally when Q value is between 0.3~0.7, community structure in network is more significant. In section 4, the definition of Q is modified to adapt to weighted network.

IV. Community Discovering Algorithm

In this section we present the event network based community discovering algorithm (ENCDA). The main idea of ENCDA is: For each edge in event network, calculate the similarity of two nodes linked by an edge, and set it as edge weight. Then apply the hierarchical community discovering algorithm on event network. At first, assume every node is a community, for any neighboring node i and j, calculate the modularity increment ΔQ when put node i into the community that node j belongs to. For every neighboring node of node i, calculate the modularity increment, and pick the largest one, if it is a positive value, then join node i into the community that the corresponding neighboring node belongs to; otherwise, node i remains unchanged. Repeat this process, until there is no change. So the first community layer is created. Now build up a new network, all of its node is the communities found in the first pass, and edge weight of the new network is the sum of all the weights of connected edges between two communities. Apply the community discovering algorithm on this layer again and get the second layer of communities. Continue this progress, until the network can't aggregate to a new network.

A. Similarity of Events

Event similarity refers to the amount of information shared between two different events. Our event similarity calculation algorithm is based on the following idea: if the corresponding event elements of two events are similar, then these two events are similar. Action element is the core element of an event. Event classification should be based on the principle that taking action element as the core consideration, and similarity of events of the same class is greater than that of different class. In other words, event elements' similarity is the measure of event similarity, the calculation of event similarity shall be turned into the calculation of each event element's similarity, and then calculate their weighted composite value.

The similarity of event element includes two parts: syntax similarity and semantic similarity. Syntax similarity means the similarity of text characters. When two elements share some common words, they are somewhat similar. Semantic similarity means whether two elements have a deeper connection in semantic level.

1) Syntax Similarity

The syntax similarity is mainly considering from the point of view of the event element's value. Event element's value is generally composed by string, so in this paper we naturally use the similarity determination method of string to calculate the syntax similarity.

For two factors l_1 and l_2 , their syntax similarity $S_{syn}(f_1,f_2)$ is:

$$S_{syn}(f_1,f_2) = \frac{2Common(f_1,f_2)}{Len(f_1) + Len(f_2)}$$

in which $Common(f_1,f_2)$ is the number of common word in f_1 and f_2. $Len(f)$ is the length of f .

Syntax similarity is an indispensable part of events' similarity. Considering the following two sentences:

a) "*A butterfly flaps its wings in South American could affect the weather in Texas.*"
b) "*The butterfly effect is a term used in chaos theory that describes how small changes to a seemingly unrelated thing or condition can affect large, complex systems.*"

For the event in sentence (a), its Action is "*flap*", Object is "*butterfly*", Environment is "*South American*". For the event in sentence (b), its Action is "*describe*", Object is "*butterfly effect*". They have no element that is semantically similar, so they are irrelevant in semantic level. But in fact they are telling about the same thing - the butterfly effect. Although the Object of these two events has no semantic relation, they have a common word "*butterfly*", so by considering their syntax similarity, these two events are related.

2) Semantic Similarity

Semantic similarity calculation requires some semantic knowledge resources as basis. For Chinese language, HowNet is widely used as a semantic knowledge source. The description object of HowNet is concepts represented by Chinese and English words. It reveals the relation of different concepts and relation between concept and its properties. A semantic similarity calculation method based on HowNet is proposed in[12].

In this paper, the semantic similarity of two concepts f_1, f_2 is defined as $S_{sem}(f_1,f_2)$.

a) If concept f_1 and f_2 are synonymous in HowNet, then $S_{sem}(f_1,f_2)=1$. For example, if the Object of event e_m is “aircraft”, the Object of event e_n is “aeroplane”, then their similarity $S_{sem}(O_m,O_n)=1$.

b) If concept f_1 and f_2 has inclusion or hyponymy relation in HowNet, then $S_{sem}(f_1,f_2)=0.5$. For example, if the Action of event e_m is “traffic accident”, the Action of event e_n is “pileup”, “pileup” is contained in “traffic accident”, their similarity $S_{sem}(A_m,A_n)=0.5$.

c) Otherwise, $S_{sem}(f_1,f_2)=0$.

3) Event Similarity

Event similarity is defined as:

$$Sim(e_m,e_n)=\sum_{i=1}^{4}(w_{syn-i}\,S_{syn-i}(f_{im},f_{in})$$
$$+w_{sem-i}\,S_{sem-i}(f_{im},f_{in})),f_i\in\{A,O,V,T\}$$

f_{im} means the i-th factor in event e_m and f_{in} means the i-th factor in event e_n. $S_{syn-i}(f_{im},f_{in})$ is their syntax similarity, $S_{sem-i}(f_{im},f_{in})$ is their semantic similarity. Since each factor has different ability in describing an event, the w_{syn} and w_{sem} of A,O,V,T are set to (0.1, 0.3), (0.05, 0.2), (0.05, 0.15), (0.05, 0.1), according to our experiment results.

B. Modularity

The definition of Q function in section 3 is modified to adapt to weighted network:

$$Q=\frac{1}{2W}\sum_{i,j}^{n}(w_{ij}-\frac{s_is_j}{2W})\delta(C_i,C_j)$$

W is the sum of all edge weights in the network. s_i is the sum of weight of all the edges that are connected to node i. w_{ij} is the weight of the edge which connect node i and j.

C. Algorithm Steps

The steps of ENCDA are described in table 1:

TABLE 1. STEPS OF ENCDA

Algorithm: ENCDA

Step 1. Calculate the similarity of two event node linked by an edge, set the similarity value as edge weight.

Step 2. Assume every node is a community.

Step 3. For any neighboring node i and node j, calculate the modularity increment ΔQ when put node i into the community that node j belongs to.

Step 4. For every neighboring node of node i, calculate the modularity increment, and pick the largest one, if it is a positive value, then join node i into the community that the corresponding neighboring node belongs to; otherwise, node i remains unchanged.

Step 5. Repeat step 3~4, until there is no change.

Step 6. Build a new network whose nodes are now the communities found in step 5. The weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities

Step 7. Repeat step 2~6, until the network can’t aggregate to a new network.

ENCDA has several advantages. Its steps are easy to implement, and the output result is unsupervised. Meanwhile it’s fast because it is easy to compute the possible gains in modularity, and the number of communities declined sharply after several loops. Also the resolution limit problem of modularity seems to be circumvented benefited from the intrinsic multi-level nature of this algorithm.

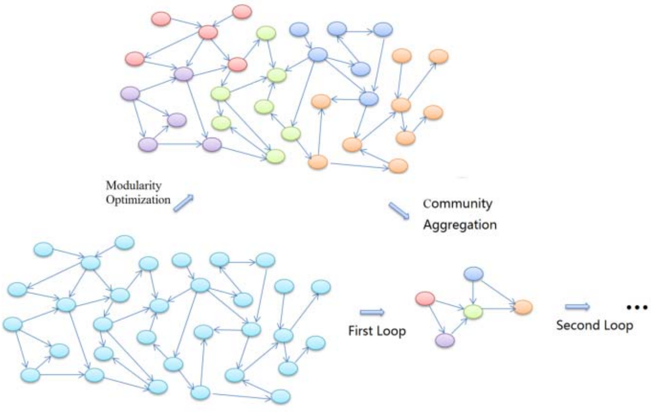

Figure 2. A schematic view of partition process

V. EXPERIMENT AND RESULT ANALYSIS

In order to verify the effectiveness of ENCDA, we use CEC(Chinese Emergency Corpus)[13] as our experiment corpus. CEC corpus is a set of emergency news articles collected from the Internet. There are five types of emergencies in CEC including earthquake, traffic accident, terror attack, bromatoxism and blaze. Event, event factors and event relations have been labelled already in each text.

In our experiments, for each type of emergencies, we choose 10 articles. First, establish event network and complete topic partition manually, then compared with partition result using ENCDA and EN-MST[2]. Evaluation indicators are accuracy(P), recall rate(R) and F value. Assume the manually divided topic set is m, the automatically divided topic set using partition algorithm is n, then define $P=(m\cap n)/n$, $R=(m\cap n)/m$, $F=2RP/(R+P)$.

Experiment results are shown in table 2 and table 3.

TABLE 2. COMMUNITY NUMBER AND EVERAGE SIZE

Type	Manually		EN-MST		ENCDA	
	C	S	C	S	C	S
Earthquake	5.1	7.3	2.9	12.8	3.7	10.0
Traffic accident	4.7	6.1	2.7	10.7	3.4	8.5
Terror attack	6.1	5.2	3.4	9.4	4.6	7.0
Bromatoxism	4.6	7.1	2.8	11.8	3.5	9.4
Blaze	5.6	4.6	3.1	0.84	4.3	6.0
Average	5.2	6.1	3.0	9.1	3.9	8.2

TABLE 3. COMPARISON OF PARTITION RESULT.

Type	EN-MST			ENCDA		
	P	R	F	P	R	F
Earthquake	1.0	0.56	0.71	0.98	0.72	0.83
Traffic accident	1.0	0.57	0.72	0.96	0.71	0.82
Terror attack	1.0	0.55	0.71	0.96	0.73	0.83
Bromatoxism	1.0	0.60	0.75	0.98	0.75	0.85
Blaze	1.0	0.55	0.71	0.94	0.74	0.83
Average	1.0	0.56	0.72	0.94	0.73	0.83

In table 2, C is the average community number in each article, S is the average community size, that is the average event number. In table 3, accuracy of EN-MST is always 100 percent, because EN-MST is a top-down division algorithm, every node will be categorized into a community, therefore no single node will left. While ENCDA is an aggregation algorithm, there may have single node which is not suitable to be treated as a topic left after the algorithm is done.

From the experiment results in table 2 we can see that size of communities obtained by using ENCDA is smaller than that obtained by using EN-MST. This is mainly because in EN-MST, the edge weight which represents the event relation is determined by the appearance frequency of event pair, which is unstable and probably influenced by the subject, content, quantity and many other factors of the articles chosen for event pair statistics. While in ENCDA, edge weight is calculated by considering the similarity of event elements, which will be more stable and reasonable. Events that are semantically similar are more likely to gather together using ENCDA, accordingly generating fine granularity communities. Results in table 3 also points to the same conclusion. In table 3, the recall rate and the quality of community partition result (F value) of ENCDA are higher then that of EN-MST.

To get higher recall rate and partition quality is not the final purpose of our research. In fact, after we finish topic detection, another important task is topic tracking. With news topic continuously growing over time, community partition results should be updated now and then. EN-MST is not able to support this task, each time to update topics it has to start over and perform community partition on the entire newly built event network. But this problem can be solved by our method. By construction, ENCDA formed a hierarchical community structure for the network, each level of which being given by the intermediate partitions found at each pass. Before the algorithm executed, communities are single node. Then in each pass of our algorithm, communities are larger and larger node sets. So when we need to update our partition result with newly joined event nodes, we can assume those new event nodes as communities, put them together with previous discovered communities, and apply ENCDA in the same way. Certainly there's still some problem to deal with in this process, for example, how to link the newly joined event nodes with previously existing event nodes. Those will be studied in our future work.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced a hierarchical aggregation algorithm to discover communities of unprecedented size in event network. The experiment results show that our algorithm can discover fine granularity communities that are semantically meaningful. It gets higher recall rate and partition quality compared to EN-MST. In our further study, the event similarity calculation method can be optimized to improve the accuracy, and we'll make our effort on topic tracking based on our work in this paper.

ACKNOWLEDGMENT

This paper is supported by the Internal Natural Science Foundation of China, No. 61305053 and No. 61273328, and Shanghai Natural Science Foundation, No. 12ZR1410900.

REFERENCES

[1] J. Allan, Introduction to topic detection and tracking, Topic detection and tracking, pp. 1-16, 2002.

[2] D. Wang, A Sub-topic Partition Method based on Event Network, ICIW 2012, The Seventh International Conference on Internet and Web Applications and Services, 2012, pp. 194-199.

[3] Kernighan B W, Lin S, An efficient heuristic procedure for partitioning graphs[J] , Bell System Technical Journal, 1970, 49(2):291-307

[4] Fiedler M, Algebraic connectivity of graphs[J], Czech Math J,1973,23(98):298-305.

[5] Pothen A, Simon H, Liou K2P, Partitioning sparse matrices with eigenvectors of graphs[J] , SIAM J Matrix Anal Appil, 1990, 11(3):430-452.

[6] Palla G, Derenyi I, Farkas I, Vicsek T, Uncovering the overlapping community structure of complex networks in nature and society, Nature, 2005, 435(7043):814-818

[7] Girvan. M, M. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):7821P.

[8] Newman M, Detecting community structure in networks, The European Physical Journal B-Condensed Matter and Complex Systems, 2004. 38(2):321-330P.

[9] Clauset A, Newman M E J, Moore C, Finding community structure in very large networks, Physical Review E, 2004, 70(6):066111

[10] Blondel V D, Guillaume J L, Lambiotte R, Fast unfolding of communities in large networks [J] , Journal of Statistical Mechanics : Theory and Experiment, 2008:PI0008.

[11] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E, vol. 69, p. 026113, 2004

[12] Qun Liu, Sujian Li, Word Similarity Computing Based on How-net, Computational Linguistics and Chinese Language Processing，Vol.7, No.2, August 2002, pp.59-76

[13] J. Fu. Research on Event-Oriented Knowledge Processing [D]. Shanghai:Shanghai University, 2010.