An Event Detection Method Based on Association Link Network

Lin Sun¹, Weijun Yang² and Xinhuai Tang¹*

¹ School of Software, Shanghai Jiao Tong University, Shanghai, China
² School of Software, Shanghai Jiao Tong University, Shanghai, China
* tang-xh@cs.sjtu.edu.cn

Abstract. With the explosive growth of network data, it becomes more and more difficult to acquire and understand information quickly and accurately. Therefore, events discovery from a large amount of document is a very challenging problem. In this paper, we propose a method to detect events, a semantic community network is constructed by ALN (Association Link Network) firstly, and then we update the word node cluster based on WCC metric, and finally merge the communities by word co-occurrence similarity. The experimental results show that this method can complete the event detection well, and it can ensure that every document in an event cluster is describing the same topic.

1. Introduction

With the progress of network infrastructure technology, the total amount of information in the network also shows explosive growth. It becomes very difficult to detect events quickly and accurately from the huge and scattered information. This paper focuses on documents clustering to represent the event semantic community. At present, there are many and mature researches on clustering methods, such as spectral analysis and information theory. However, in the field of text clustering, the applicability and effectiveness of these algorithms need to be verified experimentally. Because most text-oriented clustering methods are expanded from the representation of spatial features of word sparse distribution, but has poor clustering effect when texts at a large number. The semantics of an event is distributed in the chronologically ordered documents, and traditional approaches pay less attention on the word association links, which have a higher ability to organize semantically related words together presenting an event’s semantic information [1].

Our method focuses on news documents, which are simple and stable in structure, clear in meaning, and often contains specific phrases. Therefore, words used in the documents describing the same event are basically the same, therefore, in this paper, we construct association link network based on the frequency of co-occurrence words between documents. Based on the modified WCC metric, nodes updating strategy was proposed to divide the semantic network into communities. At the same time, community similarity is applied to implement integration between communities. Finally, documents are mapped to the community to enrich semantic network. And this network can provide support for the event tracking according to the similarity of the document and the semantic network. In the experimental part, we collect news documents about actual events from news websites through web crawlers, measure the effect of the event detection method by comparing the rate of event recognition and the degree of cohesion within the event community.
2. Related work
Several event detection methods have been proposed based on clustering algorithms. Li et al. [2] develop a topic detection prototype system based on K-means algorithm. Garcia et al. [3] proposed two clustering algorithms aiming to construct a cluster hierarchy, dealing with dynamic data sets. Chen et al. [4] present an effective Fuzzy Frequent Itemset-Based Hierarchical Clustering approach, which uses fuzzy association rule mining algorithm. Tu and Seng [5] propose a more precise set of prediction indices based on time, volume, frequency for emerging topic detection.

The common problem with these methods is that they ignore the association link between words, which can largely represent the semantic information of events. Therefore, this paper constructs ALN (Association Link Network) network through co-occurrence relationship to emphasize semantic relationship between words, so as to improve the effect of event detection.

Association Link Network [6] is a kind of semantic link network to organize the associated resources (e.g., keywords and web pages) loosely distributed in the Web, aiming at extending the loosely connected network without semantics to an association-rich network. ALN can be donated by:

\[ ALN = \langle N, L \rangle \]  

where \( N \) is a set of web resources, \( L \) is a set of weighted semantic links.

Many researchers proposed clustering algorithms based on the semantic network. Yang Liu et al. [1] proposed ALN-based event discovery approach by label propagation. But nodes in the network cannot be shared by different communities, which may lead to different event communities being divided together, or some communities miss some key words. Arnau et al. [7] propose a metric called WCC guarantees communities with structure and cohesion. WCC calculate only the number of triangles formed by current node and all other nodes in that community network, so by this metric every edge between word is the same without using weight of edges. To solve the above problems, we introduce a modified metric based on WCC which calculate perimeter of the triangle, and a node update strategy to update community partition.

3. Event detection

3.1. Preprocessing
The preprocessing part refers to the operation of word segmentation and filter. In the news text, some words in the statement are enough to clearly express the event semantics. Building semantic network with word segmentation results will lead to a large scale of the network and increase the complexity of subsequent operations. Therefore, filtering the text participle results before constructing the ALN semantic network is necessary. Filtering aims to retain key semantics in a statement (e.g., nouns, gerunds, proper nouns, and phrases, etc.). This kind of word is often found in the same event and can clearly describe the characteristics of the event.

3.2. Building ALN
For the news text, we take the word after preprocessing as the node, and use the co-occurrence frequency between words as the edge to construct the relational semantic network. Weight calculation formula of edges between nodes is calculated as follows:

\[ w(a,b) = \frac{Co(a,b)}{(DF(a) \cdot DF(b))^{1/2}} \]  

where \( Co(a,b) \) donates the times that word \( a \) and word \( b \) appear together in the same text, \( DF(a) \) donates the number of texts containing the word \( a \).

At this point, we have the original semantic network of ALN. In this network, the word nodes in the same text are all joined. Different texts that describe the same event will contain the same word combination to a large extent, so the full join graph of different texts that describe the same event will overlap, with a high degree of connection probability between the words and a high edge weight, thus
forming a highly cohesive community. Words that describe different events are less likely to connect and have lower edge weights, which can be dispersed among different communities. This is the main reason why ALN can be used for event clustering.

In order to make full use of the highly cohesive feature of ALN semantic network, after the construction of the original network is completed, an edge reduction is carried out to reduce the impact of the insignificant edge on the community division.

3.3. **Node update strategy**

In this paper, we mainly consider the perimeter of the triangle formed by current node and all nodes in that community network since the weight of edge is introduced. The difference is that WCC calculate only the number of triangles formed by current node and all other nodes in that community network. Given a semantic network $G = (V,E)$, the problem is to divide nodes of the graph into disjoint cohesive sets. We apply improved metric based on WCC to initiate community segmentation of nodes in the community. In term of a node, the degree of cohesion between the node and the community can be donated by:

$$CC(a) = \frac{1}{N_a} \sum (w(a,b) + w(a,c))$$

where $N_a$ donates the number of triangles formed by node $a$ and all other nodes in that community network, $\bar{\text{Weight}}_a$ donates the average weight of all the edges derived from $a$.

For a community $C$, the improved WCC metric are calculated as follows:

$$wcc_p(x,C) = \begin{cases} \frac{p(x,C)}{p(x,V)} & \text{if } p(x,C) > 0 \\ 0 & \text{if } p(x,C) = 0 \end{cases}$$

where $p(x,C)$ donates the perimeter of the triangle formed by current node and all nodes in the community network, $vt(x,V)$ donates the number of nodes in $V$ that can form triangle with node $x$.

For a partition of $V$ donated by $P = \{C_1, \ldots, C_n\}$ which is a pairwise disjoint subsets of $V$. For a partition $P$, the improved WCC metric are calculated as follows:

$$\text{WCC}_P = \sum_{i=1}^{n} \sum_{x \in C_i} \text{WCC}_P(x,C_i)$$

Thus, in the subsequent iterative update process, the updating strategy of community clustering is determined according to the numerical changes of the metric. In this paper, we propose four corresponding updating strategies for nodes.

**Keep.** Node $x$ still belongs to the current community, and the segmentation results of ALN semantic network remain unchanged.

$$\Delta_k = 0$$

**Detachment.** Node detach from current community $C_k$. Current partition of $V$ is $P = \{C_1, \ldots, C_k, \ldots, C_n\}$, the partition after detachment is $P' = \{C_1, \ldots, C_{k} \cup \{x\}, \ldots, C_n\}$, and $C_{k} = C_{k} \cup \{x\}$,

$$\Delta_k = \text{WCC}_P - \text{WCC}_{P'}$$

**Transfer.** Node $x$ transfer from the community $C_p$ to which it belongs to another community $C_q$.

$$\Delta_T = \Delta_D(x,C_p) - \Delta_D(x,C_q)$$

**Join.** Node $x$ joins a new community but is not detached from the current community.

$$\Delta_J = -\Delta_D(x,C_p)$$
In each iteration calculation of the community partition update, the strategy with the maximum value is selected to operate on the node. Among them, the last strategy is the key to solve the problem that the node is shared by multiple communities, and it is also the basis for subsequent community integration.

3.4. Community integration
Node update strategy indicates that the number of communities does not decline significantly in each update iteration, nodes only alternate between different communities. However, the existence of the last strategy means that the overlap between communities keeps increasing. When the overlap reaches a certain level, we can consider that the events described by these two communities are the same, so the two communities can be merged.

Based on this idea, this paper calculates the weight proportion of common nodes in two communities respectively and then calculates their cosine similarity. Two communities of similar proportions can be merged into one. It is donated by:

\[
\sum_{k=1}^{m} q_k i_k \cdot \sum_{k=1}^{m} q_k i_k C_{nw} C_{nw} C_{nw} C_{nw} C_{sim}
\]

Since the similarity judgment in this paper only considers the common nodes of two communities, small number of common nodes which proportion is similar will lead to the integration of less relevant communities. Therefore, our method introduces a threshold to exclude the effect of this case by calculating the ratio of the number of common nodes to the node number of communities with fewer nodes.

3.5. Map documents to communities
For the communities that have been clustered, our method maps the documents to the semantic community that is consistent with the same event, and then rebuild the semantic community by combining the existing community and documents to enrich the ALN-based semantic network. Our method applies mutual information to measure the semantically association strength between document \(d\) and community \(C_p\):

\[
I(d, C_p) = p(d, C_p) \cdot \log(p(d, C_p) / p(d) p(C_p))
\]

where \(p(d)\) donates the probability that document \(d\) appears. \(p(C_p)\) donates the probability that community \(C_p\) is selected is the ratio of the total weight of the internal edge of community \(C_p\) to the sum of the total weight of the internal edge of the community to be selected. \(p(d, C_p)\) is the joint probability between document \(d\) and community \(C_p\), which is composed of similarity and correlation.

4. Experiments
4.1. Data set
The clustering method proposed in this paper aims to describe the same event in each clustering text, so the experimental data set should be clustered based on the event. Therefore, we crawl news about 10 events from the web page of Tencent news website as the data set of this experiment. In the meantime, our method applies word segmentation tool to divide crawled documents and then filter out words that are not descriptive and noun.

4.2. Evaluation criteria
In this paper, recognition rate and purity are used to quantify the effect of our clustering method. The recognition rate refers to the ratio of the number of events divided by our method (donated by \(count_t\) )
to the number of actual events (donated by \( \text{count}_e \)), and the similarity between the two values indicates whether the result of our method's event recognition is similar to the actual situation.

\[
\text{recognition rate} = \frac{\text{count}_e}{\text{count}_c}
\]  

(12)

Purity is used to measure whether word nodes in a community network describe the same event, and high purity means that the event semantic community is highly cohesive:

\[
purity = \sum_{e \in F} |C_e| \cdot \text{Max}(F \_ \text{Measure}(C_e, Ev)) / P
\]  

(13)

4.3. Experimental Results Analysis

In this paper, two data sets of different quantities are selected as input for experiment and comparison with the label propagation method based on ALN mentioned above.

Table 1. Experimental results.

| Method                  | 6 actual events | 10 actual events |
|-------------------------|-----------------|------------------|
|                         | Recognition rate| Purity           |
| ALN with label propagation | 0.83            | 0.78             |
| Our method              | 1.33            | 0.85             |
|                         | 0.8             | 0.73             |
|                         | 1.3             | 0.83             |

As shown in the above table, the number of recognized events by label propagation method is closer to the actual number of events than the clustering algorithm proposed in this paper, but the clustering results are generally less than the actual number, leading to the decline in the purity of the clustering results. Since our method introduces weight on the basis of WCC and emphasizes more on the correlation between semantic nodes in community network, so the performance on purity is much better than that of the former. As the words in the same document are fully connected in the ALN semantics network, the degree of cohesion of subnets is high in the processing procedure, which leads to the less possibility of selecting \( \text{join} \) policy in the nodes updating iteration. In addition, other methods, such as the adjusted cosine similarity and the cosine similarity of both communities, are also considered in the community integration, however, the co-occurrence word cosine similarity used in this paper is more effective. Although the method proposed in this paper divide the event semantic community in slightly larger numbers than it actually is, it ensures the quality of the community in terms of purity. In other words, an event is likely to be divided into more specific communities, which is acceptable in engineering and can be corrected by other means.

5. Conclusion

In this paper we propose an event detection algorithm based on Association Link Network. This algorithm firstly constructs ALN network according to the co-occurrence relationship of words in the documents, and then adopts corresponding nodes update strategy based on WCC metric for each node, and finally coordinates with the community integration and text mapping method to achieve the event detection. The experimental results show that this method can complete the event detection well, and it can ensure that every document in an event cluster is describing the same topic. In the future studies, time dimension can be considered to be added for word nodes, and the calculation of inter-community correlation in the community integration mechanism can further improve the effect of event detection of this method.

References

[1] Liu Y, Luo X, Xuan J. Online hot event discovery based on Association Link Network[J]. Concurrency & Computation Practice & Experience, 2015, 27(15):4001-4014.
[2] Li S, Lv X, Wang T, et al. The key technology of topic detection based on K-means[C]// International Conference on Future Information Technology & Management Engineering. 2010.

[3] Reynaldo Gil-Garcia, Pons-Porrata A. Dynamic hierarchical algorithms for document clustering[J]. Pattern Recognition Letters, 2010, 31(6):469-477.

[4] Chen C L, Tseng F S C, Liang T. Mining fuzzy frequent itemsets for hierarchical document clustering[J]. Information Processing & Management, 2010, 46(2):193-211.

[5] Yang H C, Lee C H. A novel self-organizing map algorithm for text mining[C]// International Conference on System Science & Engineering. IEEE, 2010.

[6] Luo X, Xu Z, Yu J, et al. Building Association Link Network for Semantic Link on Web Resources[J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(3):482-494.

[7] PratPérez, Arnau, Dominguezsal D, Brunat J M, et al. Shaping Communities out of Triangles[J]. Computer Science, 2012:1677-1681