Research on Topic Mining Algorithm Based on Deep Learning Extension

Fuqiang Yang, Xuechen Zhao, Maohong Zhang

School of Data Science and Computer Science, Shandong Women’s University JiNan, China

Abstract: This paper proposes a new algorithm that can consider both keywords and time of occurrence. Firstly, after the data is preprocessed, the LDA model of the theme event set is established, and the set of subject words is generated as the description mark of the event. The semantic and temporal similarity between the event keywords are calculated by the DTW algorithm to obtain the corresponding similarity matrix. Finally, the collaborative training method is used to iteratively generate the final feature vector and complete the event selection. The simulation results show that the proposed algorithm has higher accuracy and higher efficiency than previous algorithms.

1. Introduction

The keyword method obtains the event class cluster by obtaining the key period in the text and its corresponding word vector, and then expanding the keyword and word vector [1]. In recent years, event extraction methods based on deep learning extension are increasingly favored by topic event capture tasks [2]. MABED also did not fully consider the impact of timing relationships on subject matter extraction. DECoW [3] fully considers the role of time series information. By composing the fluctuation entropy to form the final event by composing the keyword occurrence period, the algorithm only considers the timing information, which causes multiple events occurring at the same time to be accurate. Extracted [4-5].

Based on the previous work, this paper proposes a new algorithm that can consider keywords and time of occurrence at the same time. The core flow of the algorithm is as follows: After the data is preprocessed, the LDA model of the theme event set is established, and the set of subject words is generated. As the description mark of the event, the semantic and temporal similarity between the event keywords are calculated by DTW algorithm, and the corresponding similarity matrix is obtained. Finally, the cooperative training method is used to iteratively generate the final feature vector and complete the event selection.

2. Deep learning extension framework

The deep learning-based theme event extraction method proposed in this paper can be constructed based on the following three reasonable assumptions: (1) An event can be represented by several keywords. For example, the Paris terrorist attack on November 13, 2015 can be used in France. Keywords such as the capital, "bomb" are described. (2) A certain algorithm can be used to extract event-related keywords from topics and words. For example, "terrorists" and "bombs" are generally used to describe events such as terrorist attacks. (3) If several keywords often appear together, and the timing trends are roughly the same, then these words are more likely to describe the same event. The latter two assumptions are the most important of the current popular time extraction methods and
models. The core idea is that the subject text information and time series information have a significant impact on the extraction of events.

Based on the above assumptions, the deep learning extended event extraction method proposed in this paper uses both topic and timing information. It is mainly composed of three parts: First, the text is processed. The main processes include labeling part of speech, labeling time, filtering low frequency words, and so on. Then, a number of key topic words are extracted as candidate event feature words through the theme, and then the keywords and correlation degrees between each pair of feature words are calculated, thereby constructing a semantic similarity matrix and a temporal similarity matrix corresponding to all the feature words. Finally, input them into the deep learning extension algorithm to get a result that contains the most accurate set of keywords that can describe an event.

1) Theme model

The topic model LDA (latent dirichlet allocation) is a classic probability topic model. LDA believes that each article does not belong to only one topic, but belongs to a group of multiple topics. At the same time, the words in each article appear on a certain basis and with a certain probability. This article first builds a topic LDA model, each of which is a mixed distribution of events, and the probability of occurrence of a word in a topic is a polynomial about the event. Through the Gibbs sampling, LDA can obtain the words with high probability corresponding to each event, which is the characteristic word of this event.

The method of extracting the subject words uses the letter K to indicate the total number of topics, and the distribution of the words of the kth subject is recorded as \( \phi_k \), and the value of \( \phi_k \) can be obtained by the subject model LDA using the Gibbs sampling method. For a topic k, the m words with the highest \( \phi_k \) value are determined as the subject words. Under the premise of allowing repetition, at most k× m words will become topic feature words. The keywords selected by the above method are of great significance for the accurate description of the event and will also be used to describe the event.

Keyword similarity calculation When using the LDA model to extract topic events, each topic is usually represented as a topic phrase. These keywords can be used to describe an event, but a topic contains several events that belong to the same topic. The possibilities are great. For example, a state visit without the leadership of a country is a subject, but belongs to several events. Therefore, the results obtained by the above methods do not have practical use value. In order to enhance the usability of the results, this paper introduces a new word clustering method, which re-clusters through the topic distribution of words. The similarity of feature words in the clustering process is calculated by Cosine algorithm.

After obtaining the distribution matrix \( \phi \) of the word in the article about the subject, the keyword i can calculate the similarity of the keyword by the cosine similarity function of the formula (1) using the subject vector \( \phi_i \).

\[
\text{Sim}_s(x, y) = \text{sim}_\cos(\phi_x, \phi_y)
\]  

(1)

2) Timing analysis

EDCoW uses time series information for time extraction, which describes the change of words in the data by constructing the frequency trend curve of each keyword in the subject. The core assumption of EDCoW is that the keywords with the same frequency development trend are more likely to describe the same event. Based on the above theoretical methods, this paper defines a new calculation formula to calculate the similarity.

Key words timing signal similarity calculation The method of calculating the signal trend of words in this paper is as follows: firstly, the extracted subject data set C is time-segmented, assuming that data C is a T-day data set, and will be used as an event node every day, then the word w The signals can be listed as follows:

\[
S_w = [s_w(1), s_w(2), \ldots, s_w(T)]
\]  

(2)

In order to improve the computational efficiency and accuracy of timing similarity, this paper uses Dynamic Time Warping (DTW) to calculate the similarity of \( S_v, S_y \) two timing signals. The DTW algorithm is a classical dynamic programming algorithm that can be used to calculate the similarity
between sequences. The DTW algorithm divides a large problem into several small problems. The calculation of each small problem is based on the obtained calculation results.

(3) Deep learning extension

Although the methods in the above two chapters can complete the extraction work, there are problems respectively: only part of the data information is used. In order to make full use of various effective information, this paper proposes a variety of extraction algorithms to solve. The algorithm is suitable for the sample size to be directly decomposed, while describing the samples do not interact with each other. The deep excavation extended theme mining algorithm flow of this paper is as follows:

Constructing the keyword similarity matrix Firstly, it only needs to construct the N most important words that are rich in meaning, so that on the one hand, the event can be described more accurately, and the time and system overhead of the algorithm can be greatly reduced.

Through formula (1), we can calculate the semantic similarity between keywords, calculate the complete set of keywords separately, and construct the semantic similarity matrix G between different keywords. The G matrix is a semi-definite matrix of N×N, and the matrix term G_{i,j} represents the semantic similarity of the keyword w_i and w_j.

According to the semantic similarity matrix G, the temporal similarity matrix T can be defined by the same reason, and the matrix term T_{i,j} represents the temporal similarity of the keywords w_i and w_j.

The theme excavation algorithm of deep learning extension adopts the principle of collaborative training, taking the semantic similarity matrix G of the sample and the temporal similarity matrix T as inputs, and calculating the obtained results. The core of the method is to continuously influence the result of another perspective through the similarity matrix of one perspective, that is, the semantic similarity matrix G and the temporal similarity matrix T in the text are mutually corrected. After multiple rounds of iteration, the obtained is integrated. Optimization results from two perspectives.

The flow of the above algorithm is briefly described as follows: In the i-th training process, the normal algorithms are used to calculate the matrices of the two different perspectives, and the feature vectors U_1^{i-1} and U_2^{i-1} can be obtained. In the next iteration, U_1^{i-1} and U_2^{i-1} are brought into the algorithm, and the results at this perspective can be obtained separately. Combining the initial matrices G and T, a new semantic similarity matrix S_1 can be recalculated, S_1 is obtained by U_1^{i-1} correction, and a new temporal similarity matrix S_2 can be obtained. The above process is iterated continuously, so that the two matrices interact and be modified. After the iterative process is completed, the G matrix is modified to obtain the eigenvector U_1^{iter}, and the T matrix is modified to obtain the eigenvector U_2^{iter}. This paper considers that the importance of semantic information is far greater than the time series information. Therefore, U_1^{iter} is selected as the final calculation result. This result is influenced by both semantic and temporal perspectives. The above method is named as multi-view subject extraction method DLE.

3. Experiment and result analysis

(1) Experimental setup

This paper selects two public data sets for experiments: FSD2011[2] and Event2012[3]. Each data in the dataset is marked with an event. FSD2011 is used as a small sample dataset, and 20 of the subjects with normal number of subjects are selected for experiment. Event2012 is used as a large sample dataset, in order to further analyze different methods in different sample sizes. The difference is that the data set is randomly divided into 50, 100, 150, and 200 event packets.

(2) Experimental results

The core of DLE is based on keyword extraction. This paper compares DLE with three other keyword-based extraction methods.

Table 1 shows the experimental results of the five data sets for the four algorithms. TSC and LDA are two single-view based clustering algorithms. Their experimental results are basically the same as
DLE. The results of LDA algorithm in all data sets are not as good as those of DLE. As the number of dataset events increases, the extraction results of all methods are declining, and the TSC extraction results are the most obvious. The main reason is that the core of TSC is to distinguish the keywords with different timing trends, so it cannot identify Events that occur at similar times occur. In the process of increasing the amount of data, the DLE method always maintains a high accuracy.

| Dataset Method | MABE | TSC | LDA | DLE |
|---------------|------|-----|-----|-----|
| Precision     | 0.9100 | 0.7100 | 0.9100 | 0.9100 |
| Recall        | 0.5600 | 0.6100 | 0.7100 | 0.8100 |
| F-score       | 0.6928 | 0.6561 | 0.7975 | 0.8571 |
| Precision     | 0.8433 | 0.6100 | 0.7700 | 0.8700 |
| Recall        | 0.1900 | 0.5700 | 0.7100 | 0.8700 |
| F-score       | 0.5600 | 0.6100 | 0.7100 | 0.8100 |
| Precision     | 0.8059 | 0.4600 | 0.6700 | 0.7700 |
| Recall        | 0.3600 | 0.4200 | 0.6400 | 0.7200 |
| F-score       | 0.4962 | 0.4391 | 0.6547 | 0.7441 |

Next, the DLE is compared with the text-based extraction method. LSH [2] is a typical representative of this method. The core is to use only the semantic information of the text to estimate the text similarity, the use of local sensitive hash (LSH). Can improve the efficiency of calculation. In order to enable DLE to compare with it, this paper deals with DLE by clustering keywords into key class clusters. In addition to the several data indicators mentioned in the previous section, this section uses the Normalized Mutual Information (NMI) indicator for evaluation. The experimental results are shown in Table 2.

| Dataset Method | FSD2011 | Event2012_1 | Event2012_2 |
|---------------|---------|-------------|-------------|
| Precision     | 0.2079  | 0.0579      | 0.0427      |
| Recall        | 0.9440  | 0.8187      | 0.8278      |
| F-score       | 0.3365  | 0.8187      | 0.8278      |
| NMI           | 0.7280  | 0.5817      | 0.5661      |
| Precision     | 0.8441  | 0.8581      | 0.8208      |
| Recall        | 0.8417  | 0.7380      | 0.5914      |
| F-score       | 0.9072  | 0.8762      | 0.8486      |

Table 2 NMI comparison of different algorithms

| Dataset Method | FSD2011 | Event2012_1 | Event2012_2 |
|---------------|---------|-------------|-------------|
| Precision     | 0.2079  | 0.0579      | 0.0427      |
| Recall        | 0.9440  | 0.8187      | 0.8278      |
| F-score       | 0.3365  | 0.8187      | 0.8278      |
| NMI           | 0.7280  | 0.5817      | 0.5661      |
| Precision     | 0.8441  | 0.8581      | 0.8208      |
| Recall        | 0.8417  | 0.7380      | 0.5914      |
| F-score       | 0.9072  | 0.8762      | 0.8486      |

4. Summary
This paper proposes a topic extraction task based on Deep Learning Extension (DLE) and constructs multiple frameworks to simultaneously utilize topic information and timing information. The comparison of DLE with other similar algorithms proves that DLE has significant efficiency and accuracy.

References
[1] Sarma A D, Jain A, Yu C. Dynamic relationship and event discovery[C]// ACM International Conference on Web Search and Data Mining. ACM, 2011:207-216.
[2] Feng L, Peng W, Zhu E. Efficient Grid-based Clustering Algorithm with Leaping Search and Merge Neighbors Method[J]. 2017.
[3] Sheng X, Xin W, Luo Y. A novel text mining algorithm based on deep neural network[C]// International Conference on Inventive Computation Technologies. 2017.
[4] Li Y. Research on Hotspot Topic Discovery Algorithm Based on Web Mining Technology[C]// IEEE International Conference on Computational Science & Engineering. 2017.
[5] Wang C, Du Y J, Ming W T. Opinion leader mining algorithm in microblog platform based on topic similarity[C]// IEEE International Conference on Computer & Communications. 2017.