Research on Focus Content Tracking Method Based on Online Message Evolutionary

Dapeng Zhou¹, Yong Zhu¹

Abstract

Aiming to the data flow continued to flow into the system in the process of training, it leads to the subject of each paragraph with constant change. This paper propose a topic tracking algorithm for online message delivery. And the paper analyzes the distribution feature of the topic focus in the text, and introduces the main steps and the algorithm model of the focus recognition method based on the analysis results. The text content comes from the real data of the real network, and the method is verified by experiments, the experimental results show that the proposed method can effectively obtain the focus of the topic development process, and can be expressed in the form of key words set and statement set.

Keywords: Focus identification; topic tracking; online social topics; Topic model;

1. Introduction

When reading these network flow text data, people will find that hot topic of network flow text focus content is changing with time. For example: the network news coverage of an event is constantly updated, but the content of the report is different; Network reviews on a topic of the network are constantly being published, but the focus is gradually changing. To identify and find timely that the migration content focus network topics, can more fully understand the network topic information structure and evolution trend, for analyzing the network public opinion situation, determine the expected public opinion has very important value to research, it is an important research problem in the field of network public opinion to analyze. But in the actual data, whether it is the text of the network news or blog, forum, they have a large amount of difficult with redundant information content and the content of the feature extraction and so on, in order to identify the content of the topic of the network, it puts forward a great challenge.

The object in this paper is the text of the network flow, through the analysis of the characteristics of the focus content of the network, the paper proposes a method to identify the topic content. The content of the network topic can be caused by the hot topic of the network topic, the impact of the event on the community, the user comments on the topic, the subsequent development of related events, etc., also includes the focus on a topic to start, to reach the peak, the gradual disappearance of the evolution of the process and the impact of the relationship between the various content and so on. In this paper, we put forward a method to identify the content of the topic in the network, and then the validity of the proposed method is verified by experiments based on the actual data of the network flow.

¹Gongqing College, Nanchang University ,Gongqing City, Jiangxi, 332020, China
2. Online message propagation algorithm

In order to improve the speed and accuracy of the LDA model on line learning algorithm for processing massive data sets and data streams, Online Belief Propagation (OBP) algorithm is proposed in this paper. The main idea of this algorithm is shown in Figure 1. OBP algorithm divides the whole data set into a series of small data segments. For the first data set, the OBP algorithm is the same as the BP algorithm, and save the result parameters $\phi$ for the current section after the training. From the second section to the last section of the data set, the OBP algorithm directly uses the training result parameters of the previous paragraph and then uses the gradient descent method to calculate the current segment of the message. For each parameter $\phi$, it is updated when the convergence of the OBP algorithm or the maximum number of iterations. According to the online stochastic optimization theory [2], this paper sets the weight of the current segment and the training segment is among parameters $\kappa \in (0.5,1]$ is used to control the data set that has been processed, constant $\kappa$ is used to reduce the impact of the iteration at the beginning of each segment, the variable $t$ represents the number of segments of the current training data. In Figure 1, $M$ indicates that the total number of data sets is divided into a large number of data sets, where $M = 1$ and $\kappa = 0$, the online message passing algorithm is converted to off-line message passing algorithm.

![Figure 1. The flow of online learning algorithm of LDA model.](image)

In this paper, the data set is defined as, Subject distribution variable is $\theta$, the probability distribution of variables of Subject corresponding to the word list is $\phi$, the word hidden topic variable is the distribution.

In the process of training, the first paragraph of the text in the 0 and 1 random initialize the first paragraph of the parameter and the normalization process is carried out to ensure that the save parameters after training. From the second to the last paragraph, the OBP algorithm only needs to initialize the parameter of the current segment, the training result parameters of the previous section are parameters, and update message until convergence. In this algorithm, we can determine whether the current convergence of the segment according to the difference between the adjacent iteration of the parameter, it can also be used to determine convergent according to the difference between the adjacent iteration and the equivalent parameter. Because when updating parameters, parameters are fixed, so only are mutually influenced. Based on the convergence of the message, the estimated parameters are for the training of the results of the parameters is to take weights sum between the current segment and has been training segment: Table 1 gives the specific implementation steps of OBP algorithm.
Table 1. Online message passing algorithm.

In:
Out:
Definite:
Random initialize the parameter for the first paragraph, and normalization
For \( t=0 \) to \( \infty \)
Do
  Initialize the parameter, and normalization
Repeat

Until
  Calculate the current segment:

End for

3. Experimental results and comparative analysis

3.1 Evaluating learning parameters

Evaluation of learning parameters verify the efficiency and accuracy of the OBP algorithm in the LDA model, in this paper, experimental verification was carried out in four groups of massive data sets: blog[3] is the United States politics and political blog, enron[4] is the Usa Energy Inc Enron email, nytimes[4] is the New York Times news digest and abstract of pubmed[4], table 2 gives the specific size of the four sets of data, where \( D \) represents the total document data set contains the number. \( W \) said the data set corresponds to the total number of words in the table.

Table 2. Comparison of four data sets.

<table>
<thead>
<tr>
<th></th>
<th>Blog</th>
<th>Enron</th>
<th>Nytimes</th>
<th>Pubmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>5200</td>
<td>277101</td>
<td>310007</td>
<td>7925635</td>
</tr>
<tr>
<td>( W )</td>
<td>33522</td>
<td>41221</td>
<td>112552</td>
<td>152210</td>
</tr>
</tbody>
</table>

Three learning parameters are introduced in the OBP algorithm proposed in the LDA model in this paper: parameters \( \kappa \in (0.5,1] \) is used to control the slowness of the forgotten data segments that have been trained; constant, It is used to reduce the impact of the initial iteration of the data set; \( S \) represents the number of documents that are included in each section of the data set. The training results of the online learning algorithm of LDA model are closely related to the setting of the three parameters. Usually for the selection of \( S \) value is the bigger the better within the scope of the memory capacity. When the memory capacity exceeds the data set hours, or select the \( S = D \), the massive data sets are not segmentation, online learning algorithm is equivalent to the traditional off-line learning algorithm. The selection of three optimal parameter values in Enron and nytimes data sets, respectively, is given by Table 3 and table 4. Through the contrast analysis of the prediction of the test set, when, the obtained predictive value is minimal, And the degree of confusion reaction is the current training model to predict the unknown data, the smaller the degree of confusion, the better the predictive ability. Therefore, the experiments select, The selection of \( S \) value needs to be determined according to the size of the data set and the memory capacity.

Table 3. Comparison of different parameters for Enron corpus.

<table>
<thead>
<tr>
<th>( \kappa )</th>
<th>( \theta^0 )</th>
<th>( T^0 )</th>
<th>( S )</th>
<th>( \text{Perplexity} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>1024</td>
<td>1024</td>
<td>4</td>
<td>5510.2</td>
</tr>
<tr>
<td>0.8</td>
<td>1024</td>
<td>512</td>
<td>16</td>
<td>6323.5</td>
</tr>
<tr>
<td>0.7</td>
<td>64</td>
<td>256</td>
<td>64</td>
<td>4100.2</td>
</tr>
<tr>
<td>0.6</td>
<td>64</td>
<td>256</td>
<td>128</td>
<td>2930.6</td>
</tr>
<tr>
<td>0.5</td>
<td>64</td>
<td>256</td>
<td>256</td>
<td>2566.3</td>
</tr>
<tr>
<td>0.4</td>
<td>64</td>
<td>256</td>
<td>1024</td>
<td>2048</td>
</tr>
<tr>
<td>0.3</td>
<td>64</td>
<td>256</td>
<td>4096</td>
<td>1968.4</td>
</tr>
</tbody>
</table>
Table 4. Comparison of different parameters for nytimes corpus.

<table>
<thead>
<tr>
<th></th>
<th>0.9</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
<th>0.5</th>
<th>0.4</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
<td>512</td>
<td>256</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>$T_0$</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>256</td>
<td>1024</td>
<td>2048</td>
<td>4096</td>
</tr>
<tr>
<td>$s$</td>
<td>12755</td>
<td>14667</td>
<td>11000</td>
<td>8225.3</td>
<td>5332</td>
<td>4106.2</td>
<td>3968.6</td>
</tr>
</tbody>
</table>

3.2 Performance comparison of OBP algorithms

In this paper, we use the OBP algorithm for massive data training based on the following assumptions: after the massive data segmentation, the training of the posterior segment is dependent on the training results of the previous paragraph. In order to verify the correctness of the hypothesis, figure 2 shows a set of nytimes data set in the comparison chart, rely and independent respectively after weight training results and rely on anterior posterior segment of the training is completely independent from the preceding training results. The contrast evaluation index used in the experiment is the time consumed by the degree of confusion and training model. In the LDA model, the confusion degree is used to evaluate the accuracy of the test set according to the results obtained from the training data set training. The smaller of the confusion degree is the better the effect.

As can be seen from Figure 2, in the case of completely independent of the previous training results, not only the model's predictive confusion degree value is large, but also the training is more time consuming. This is because that if the front later adopted the training results of the model when the end of the training has retained all the training section data set contains information, so it is more accurate to predict; and includes more information, model iterations required for convergence is less, so the consumption of less time. The experimental results show that the OBP algorithm is feasible to train the LDA model for the segmentation of massive data sets.

4. Conclusions

The classification and management of massive data and flow data is a difficult and hot spot in Natural Language Processing, which is great significance in people’s daily life. However, the premise of effective classification and management of document data is to have a clear understanding of the content of the document. The research of topic model is to quantify the document classification management into the practical method that the computer can be processed. This paper studies the online learning algorithm of the topic model based on massive
data and data stream, which provides an effective solution for the existing problems such as insufficient memory and incomplete data sets.

**Reference**


