Research on System Log Analysis Based on Bidirectional LSTM

Jiantao Zhao, Ke Wang
Institute of network and information systems, North China Electric Power University, Beijing, China

Abstract System logs often record important information in the system and are an important source of information on user behavior, system abnormalities, and system operating conditions. With the stacking of system time and the surge of users, the data generated by the system has also exploded. In order to achieve intelligent system management, this paper proposes a system log analysis method based on bidirectional LSTM. The method uses neural networks to extract keywords and data vectors based on the logs generated by the system, and uses the data sequence corresponding to the keywords as input to make judgments about system abnormalities. The experimental results show that the method can effectively judge the abnormality of the system.

1. Introduction
With the vigorous development of China's Internet industry and people's increasing living standards, people are increasingly relying on the Internet, and there is no doubt that the demand for the stable operation of the system is also increasing. Often, a medium-level system can generate hundreds of thousands of system logs every day. As time goes by, the system data is more and more, and the system log is more and more complicated. In order to better adapt to the new generation of the Internet, Internet companies need to continuously improve the management level of their systems.

The system log is an important part of system management, which contains important information such as user behavior and system status, and has high analysis and mining value. Today, the logs of most systems are recorded in the form of journals, and the logs are placed on the server in the form of text. Every day the system's log system will package it and send it to the designated place. Some large commercial systems will generate a large number of system logs. Very large systems like Taobao will even classify and save the logs, such as user behavior logs and system logs. Run logs, etc. The larger the system log, the more valuable it is for commercial discovery. Through the classification and mining of these logs, you can better understand the user's behavior, and you can also know the running status of the system, which can undoubtedly maximize the use of the logs generated by the system every day.

The logs generated by large systems are often very complicated, and ensuring their normal operation is more complicated, and the manpower is limited, which also speeds up the study of system intelligent management. Analysis of the system logs generated by the system at any time can often quickly find abnormalities in the system, which is more conducive to solving system abnormalities and effectively improves the stable operation of the system.

2. Research Status
In recent years, the rapid development of machine learning and natural language processing has undoubtedly accelerated the development of automatic anomaly detection for operation and
maintenance. Both supervised learning and unsupervised learning have their own irreparable shortcomings and are incapable of working in the field of log anomaly detection [4]. Traditional anomaly detection algorithms include statistical hypothesis testing and isolation forest algorithms. Zhihua Zhou proposed the isolated forest algorithm. This method detects outliers by outliers on the sample points. However, when there are many outliers, the effect is average.

In reference [1], a cross type log exception detection mechanism, logmerge, is proposed, which uses the semantic similarity of most logs to classify and judge logs. This provides another solution for some researchers who judge system exceptions based on log system. Through this mechanism, multiple logs of large-scale system can be analyzed and studied directly.

The statistical hypothesis test in statistics also performs well in the log based anomaly detection, which is also the most easily understood algorithm. Generally speaking, it is assumed that all data are subject to a certain distribution law, and the sample mean value \( \mu \) and variance \( \sigma \) are calculated respectively. As long as the detection point is between \( \mu \pm 3\sigma \), it is judged that the point is normal, otherwise it is abnormal.

In the system abnormality detection based on log analysis, the real-time judgment of the system is very high. Therefore, this paper proposes a bidirectional LSTM system anomaly detection algorithm. First, use keyword learning to get the representation of the features in the log, and perform bidirectional LSTM training on the data sequence corresponding to the keyword, so as to realize the capture of abnormal signals generated in the log system.

3. Anomaly Detection Algorithm Based on Bidirectional LSTM System

3.1. LSTM Algorithm

Long short-term memory model (LSTM), as a recurrent neural network (RNN) with special properties, has the advantages that general RNN networks do not have, and it can make up for the defects of general RNN networks in the long-term dependence problem. The general RNN network algorithm flow is shown in Figure 1. Among them, \( X(t) \) represents the input at a certain time and \( H(t) \) is the corresponding output. S is the memory unit of the training network and the neural network. The expression formula of S is as follow.

\[
S(t) = f(U_{S(t)} + W_{S(t-1)})
\]  

(1)

LSTM was proposed by Hochreiter & Schmidhube in 1997. Because of its good performance in various fields, it was later improved by researchers. It is currently widely used in natural language processing [4]. In order to save all historical information, LSTM has a memory storage unit, which is specially used to record all records before the detection point. The difference between LSTM and RNN is that there is only one network layer in each neural unit of RNN. LSTM has four network layers.

![Figure 1 RNN Network Flow Chart](image1)

![Figure 2 Internal Diagram of LSTM Unit](image2)

The first \( \alpha \) refers to the forgetting gate. This gate is mainly used to determine what information is not needed by the model. If not, it will be discarded by these. After inputting \( h(t-1) \) and \( x(t) \), a value between 0 and 1 is output. This value indicates how much the \( D(t-1) \) model retains for
historical information in, for example: 0 means discard all and 1 means keep all. Then the expression of the corresponding output $f_i$ is: $f_i = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ (2).

The second $\alpha$ and tanh gates determine whether to update the model for the detection point, that is, to retain this information. After inputting $X_t$ and $h(t-1)$, some new features are obtained. These features may be further updated into the model. among them:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$ (3)

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$ (4)

Next, the model is updated from $D(t-1)$ to $D(t)$, and the LSTM algorithm adds a screening gate here to filter out some new model features. The updated formula is as follows:

$$D(t) = f_i \ast D(t-1) + i_t \ast C(t)$$ (5)

After updating the characteristics of the model, a sigmoid layer is needed to determine whether the detection point is abnormal. Finally, the algorithm will get a value between -1 and 1 through the tanh layer. Multiply the two to get the final output $H$ of the detection point then:

$$o(t) = \sigma(W_o[h(t-1), x(t)] + b_o)$$ (6)

$$h(t) = o(t) \ast \tanh(C_t)$$ (7)

3.2. Bidirectional LSTM

Although LSTM can store the records before the detection point, in the early stage of model training, the model may often fail due to insufficient data. In general, when training an LSTM model, there is often a complete data set. Bidirectional LSTM can avoid the problem of insufficient data in the initial model training. Bidirectional LSTM can not only use the data before the detection point to modify the model, but also make model adjustments based on the data characteristics after the detection point. Use LSTM to grab useful information forward and backward respectively.

Bidirectional LSTM consists of two LSTMs in opposite directions, training forward and backward respectively. And use Softmax to adjust the final model. Bidirectional LSTM is shown in Figure 3.

![Figure 3 Bidirectional LSTM model diagram](image)

The X is the input layer, and Y is the output layer, and S is the forward LSTM model, and S‘ is the reverse LSTM model. The two are trained separately to obtain h and h’, and finally handed over to Softmax for model adjustment.

4. Experimental Results and Analysis

4.1. Experimental design and evaluation index

Usually a system log should contain time, type, ID, and specific information. This shows that this is not structured text, so it needs to be decomposed to extract the information in it and perform anomaly detection. Figure 4 shows a log in the win10 system.
This article uses the public data set wordcount data set and cross-training, using text analysis tools to pre-process and classify the system log to obtain data such as system memory and digitize it. Then extract the features of different categories of logs through training, and finally obtain the model through training.

| Log Type | Time  | Data Size | Number of Logs | Abnormal Number |
|----------|-------|-----------|----------------|-----------------|
| BGL      | 7 months | 715 M     | 8056           | 864             |
| HDFS     | 8 hours | 400 M     | 5832           | 621             |

In the field of anomaly detection, the detection point is either normal or abnormal, so it is basically a binary classification problem. The evaluation indicators commonly used for classification problems are: accuracy rate (P), recall rate (R), and F1 value. They are calculated from the classification confusion matrix, and the calculation formula is as follows:

\[
P = \frac{TP}{TP + FP} \quad (8)
\]

\[
R = \frac{TP}{TP + FN} \quad (9)
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \quad (10)
\]

### 4.2. Results and Analysis

In order to show the difference between LSTM and bidirectional LSTM, this paper will conduct experiments on both, the experimental results are shown in Table 3.

| Index | LSTM     | Bidirectional LSTM |
|-------|----------|---------------------|
| P     | 0.8043   | 0.8046              |
| R     | 0.8409   | 0.8411              |
| F1    | 0.8221   | 0.8224              |
| Training Time | 12h | 8.5h               |

From the experimental results, we can see that the indicators of the two-way LSTM comparison and the LSTM model are slightly improved, and the time it takes for the model to converge can be shortened by about 30%. At the same time, this paper also modify the parameters of LSTM on this basis to ensure the credibility of the results. This includes the number of hidden layers, and the number of neural units in each layer.
The results of this experiment show that LSTM and bidirectional LSTM have good performance in log analysis and anomaly detection, and will not be inferior compared with the models in other papers with the same data set. Compared with the traditional LSTM model, the bidirectional LSTM has a faster convergence speed in the experiment, and can save an average of about 30% of the time.

At the same time, on the premise of ensuring accuracy, the trained two-way LSTM model parses and detects a log in about 20ms. This is enough to detect the system log generated at any time in the general operation and maintenance system. This shows that the model proposed in this paper can detect anomalies quickly and accurately in most cases.

5. Conclusion
This paper presents a research method of system log analysis based on bidirectional LSTM. This method first disassembles and analyzes the system log in order to extract the characteristics of different types of logs, and then accelerates the model convergence through the simultaneous training of forward LSTM and reverse LSTM, and finally realizes the abnormal detection of system logs. The experimental results show that the method can quickly and effectively judge the abnormality of the system, and can satisfy the automatic operation and maintenance of most small and medium-sized systems.

However, anomaly detection based on logs is only a small part of log analysis, and this article is limited to this. The log system also has user behavior logs, which have greater commercial value for its analysis and can dig out more information from it, which is also the next step of the work plan.

Acknowledgement
I am very grateful to the school for providing me with a research platform, and I am also grateful for the strong support of the laboratory.

References
[1] Zhang Shenglin, Li Dongwen, et al. General anomaly detection mechanism for multi-syntax log for cloud data center [J]. Computer Research and Development, 2020, 57 (04): 778-790 (in Chinese).
[2] Qiu Yuan, Chang Xiangmao, et al. Stream data anomaly detection method based on LSTM and sliding window [J/OL]. Computer application: 2020, 05, 28 (in Chinese).
[3] Yang Ruipeng, Qu Dan, et al. Improved log sequence anomaly detection in time convolution network [J/OL]. Computer Engineering, 2020, 05, 28 (in Chinese).
[4] Wang Haining, Yuan Xiangfeng, et al. Network traffic prediction and application based on LSTM and traditional neural network [J]. Mobile Communications, 2019, 43 (08): 37-44 (in Chinese).
[5] Zhang Shuangjiang. Design and Implementation of Intelligent Monitoring System Based on XGBoost and LSTM [D]. Nanjing University, 2019 (in Chinese).
[6] Lu Peiyao. Research and implementation of software system anomaly detection method based on LSTM [D]. Xidian University, 2019 (in Chinese).