BRAS syslog pattern generation method: a preliminary experiment on clustering algorithms

1st Li, Yun-Jie
Network Management Technical Laboratory
Telecommunication Laboratories, Chunghwa Telecom
Taoyuan, R.O.C. (Taiwan)
Emails: michael@cht.com.tw

2se Li, Jhao-Yin
Network Management Technical Laboratory
Telecommunication Laboratories, Chunghwa Telecom
Taoyuan, R.O.C. (Taiwan)
Emails: jhaoyinlee@cht.com.tw

***Abstract*—Internet service provider (ISP) uses Broadband remote access server (BRAS) to connect its customers called subscribers. Due to the usage of the Internet had grown up rapidly, the ISP company have to pay more attention to manage BRAS device. In general, the ISP administrators set up several suspicious Syslog extraction patterns into the operation support system (OSS) to extract Syslogs. The administrators will check out the detail of the Syslog matched the suspicious Syslog extraction pattern to find out if any problem exists on devices while they are notified by OSS that there are some words in Syslog are matching the extraction pattern.**

However, it is difficult to define the proper extract pattern that the administrators do not see before. Especially, after configuration had been changed, e.g. after an upgraded new version of the software or after adjusted the topology of the network, etc, there are a few new words related anomaly in Syslogs and those words are not match the Syslog extraction pattern because the administrators do not see those before.

Furthermore, we found that BRAS Syslog data have a special feature that the other log data would not have is that there is a lot of content-related digits. Note that, those contents carry unique information and can make the extraction algorithm misleading. According to that, how to leverage the automatically extract algorithm to deal with this problem is another issue we have to take care of.

In this paper, we proposed the BRAS Syslog pattern generation methodology and conduct a preliminary experiment to evaluate the performance. The result shows that the method with the combination of tfidf and EM or tfidf and HCA can produce better performance compared with the other clustering algorithms.

***Index Terms*—Machine Learning, Unsupervised Learning, Clustering algorithm, Network Diagonosis**

I. Introduction

Broadband remote access server (BRAS) device provides Internet service to the subscribers by routing the Internet resources from the ISP network to the subscribers. During the last decade, the demand for access to the Internet has grown significantly and it would increase the cost related to manage the devices. While the loading of access to the Internet grows up, the complexity and cost of management for devices grows up too. Given the fact that it is difficult for the ISP to prevent the devices from the trouble that would shutdown the device or take a negative influence on the subscribers.

In order to detect the problem and deal with it proactively, the administrators defined beforehand several log extract pattern which is used to alarm to the administrators that there is some critical or abnormal situation happened on the devices. In other word, when the administrators guess that there are some words that it is related to a problem that would take a negative effect on the device and it would not make the device shutdown immediately. They can define that words in Operation Support System (OSS) to let the system informs the administrators while the words appear in the Syslogs to remind them to check the situation by reading the detail of Syslog of BRAS.

However, it is difficult to define the log extract patterns that people don't see before. To define the patterns, the administrators have to spend a lot of time analyzing the Syslogs to recognize which word is suspicious and can be defined as the pattern to prevent the device from error proactively.

On the other hand, how to figure out the pattern that we do not see before is not just one issue. The behavior of BRAS on Syslog is the other issue we have to pay attention to. BRAS device routes traffic between the subscribers and the Internet resources that the subscribers want to access to and it would produce a lot of words that contain digits on its Syslog, such as IP address related to the host, network area-id, connected network resource of BRAS, etc. It would not be effective while we want to figure out the new log patterns. Because those words have critical information and unique identification, it would mislead the analysis if we want to use an algorithm to find out the new pattern from data automatically.

In the previous research, there are many log analysis tools like Swatch [1] that are rule-based and must be defined by domain knowledge to monitor the logs. Those tools can not be used to generate log patterns automatically because of lacking knowledge about unknown words. LogCluster [2] see the log analysis as the pattern mining problem. It can be used to extract the log pattern, but its methodology does not include the mechanism of dealing with the problem of digits words that may interference with the result of the analysis.

Because of the behavior of BRAS, there are a lot of words that contain digits to indicate the situation of connection with subscribers or the other connected network, such as the connection status indicated the ip address of subscribers or the connection status with the other network domain connected to the Internet. Those words may have a significant impact regarding the value of the support threshold that is the common

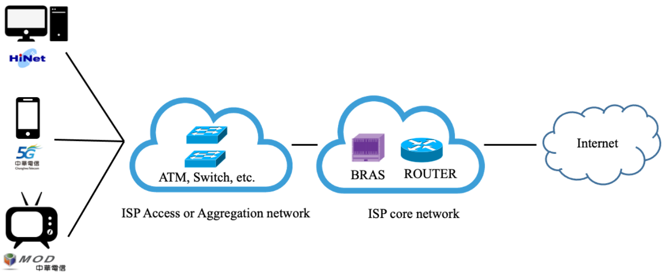

Fig. 1. An illustration of structure of ISP network

measure criteria in the theory of data mining because the words have high frequency and have unique information that makes mining algorithms see it as critical words.

Therefore, there are two main issues if we want to deal with the problem of generating suspicious Syslog extract automatically. The first one is how to find out the new pattern without domain knowledge involving. The other one is that we have to concern the question about the digits that may have a significant impact on any data analytics method while we want to extract the log automatically. How to prevent the adverse impact from those digits word is the critical problem while we want to apply automatically extraction on BRAS.

II. Overview of ISP network and operation support system

ISP company deploys BRAS on its network to provide the service that allows the user to connect Internet. Fig. 1 depicts the structure of the ISP network from a BRAS perspective. The network level of BRAS is between ISP access/aggregation network and ISP core network. The former is the network that deploys access or aggregation devices to connect and merge the customers, such as the layer 2 device, wireless cells, and so on; the latter is the network that deploys the critical devices to handle how to access the resources of the Internet for customers, such as the BRAS, Router, etc.

III. Methodology

In this work, we proposed the extraction method that merged two concepts to deal with the problem of automatic extracting log patterns. The first one is that we see the Syslog as the documents set and use tfidf (Term Frequency - Inverse document frequency) [3] to establish the word matrix. The second one is using the clustering algorithm [3] to recognize the new word pattern that the administrator does not see before.

We would like to experiment to check out whether the extraction methodology we proposed can take effect on the BRAS device or not. The detail about the extract method we proposed is that we use tfidf equation to calculate the value for each word on the Syslog as a matrix. And then, we use the clustering algorithm on the tfidf matrix and choose a cluster produced by the clustering algorithm that has a minimum of the sum of the tfidf values compared with the other clusters.

Therefore, we have interviewed the staff who are responsible for maintaining BRAS in our company to get the words list contained a few Syslog words which would take adverse effect on the BRAS judged by their recent experience. The word list is not a piece of rigorous information, i.e., that list can not be proved which word have a strong relationship with the breakdown status of the device. Instead, the word list can see as the cue that we can imply the result produced by the algorithms may have a good effect on the data by comparing with this word list.

This word list can be named ‘suspicious word list‘ because it is written by the administrators with their domain knowledge and experience at that time. Given that reason, we would like to name this criterion as ‘match ratio‘ instead of ‘accuracy‘ which is a common measure for the field of machine learning.

The evaluation criteria for the experiment is to compare the set of words produced by algorithms and the suspicious words list which we got from staff and to calculate how many words exist both and get the value of ‘match ratio‘.

A. *TF-IDF*

We use TF-IDF [4] method to vectorize the Syslog and to calculate the influent value for every word that exists in Syslog. The main concept of TF-IDF is to calculate two values relating to term frequency and the number of the document that contains the terms respectively. In this work, we see each device as a document and each word exists on the Syslog as a term. Given D as the total device set and t is the word that exists on device d. The definition of term frequency (the frequency of the words on Syslog) is:

$$tf_{t,d} = \frac{freq_{t,d}}{\sum_{t' \in d} freq_{t',d}}, \tag{1}$$

where $freq_{t,d}$ is the word counts regarding to word t on device d.

This value can give us how important the words are. But it would mislead us if we only use this value. Because the common words would provide high term frequency among the several documents. Given that bias, we have to calculate idf value to make sure that the word we pick out is important and not a common word.

The definition of inverse document frequency (the number of how many devices contains the word) is:

$$idf_{t,D} = log \frac{|D|}{1 + |d \in D : t \in d|}, \tag{2}$$

where $|D|$ is a total number of device and the denominator ”$1 + |d \in D : t \in d|$” is calculating how many device that has the word t.

Finally, those values can be used to reflect the important coefficient for each word by multiplying them and be used to establish the tfidf matrix:

$$tfidf_{t,d} = tf_{t,d} \times idf_{t,D} \tag{3}$$

B. *clustering algorithm*

The clustering algorithm is one of the machine learning theory algorithm [3] and it is used to analyze historical data to label data as several groups name. Those groups consist of one or more data points that their features are similar within the group. Because the clustering learning algorithm does not require the label data that is used to evaluate the performance of the machine learning algorithm, there are no actual answers to compare with the result produced by the clustering algorithm to check whether it is correct or not, i.e., What the effect that the clustering algorithm provides depends on the domain knowledge.

In this work, we choose eight famous clustering algorithms [5] to evaluate which algorithm can provide better results to the BRAS Syslog automatic extraction problem. The clustering algorithms are DBSCAN (Density-Based Spatial Clustering of Applications), KMeans (K-Means), AP (Affinity Propagation), MeanShift, SC (Spectral Clustering), HCA (Hierarchical Clustering Analysis), OPTICS (Ordering Points To Identify the Clustering Structure) and EM (Expectation-Maximization). Each of them has special property because of their algorithm and we want to analyze the result of combining those clustering algorithms and tfidf method.

In order to eliminate the impact of the words containing digits, we leverage the clustering learning algorithm on tfidf matrix to pick out the words which belong to a cluster that has the lowest sum of tfidf value compared with the other clusters.

IV. Experiments

In this section, we provide an introduction to the dataset used in this work and display the experiment to show that the result while we use the combination of tfidf method and clustering algorithm on the BRAS device. Those algorithms used in this work are leveraged by sklearn package [5]. As mentioned in Section III, the experiment display tfidf method combined with the clustering algorithms, i.e., each clustering analyses the tfidf matrix to extract the words that belong to a cluster with the lowest sum of tfidf value in the matrix. After we get the list of words produced by the combination of tfidf and clustering algorithm, we use it and the suspicious words list to calculate the match ratio to see each performance.

A. *Data Set*

BRAS is the critical device for the ISP company and its content contains a lot of perspectives of network for displaying plentiful information. The structure of data is well defined and that is convenient for loading into the other place for monitoring or analyzing.

Given the fact that we don't change or adjust the contents although its diversity may give an adverse effect on the clustering algorithm. We choose the BRAS device which is owned by our company located in Taipei city because it involves all types of contents produced by BRAS and it can ensure that the data we collected is comprehensive. We collected March 2021 logs and got 257693 lines to be used for this work.

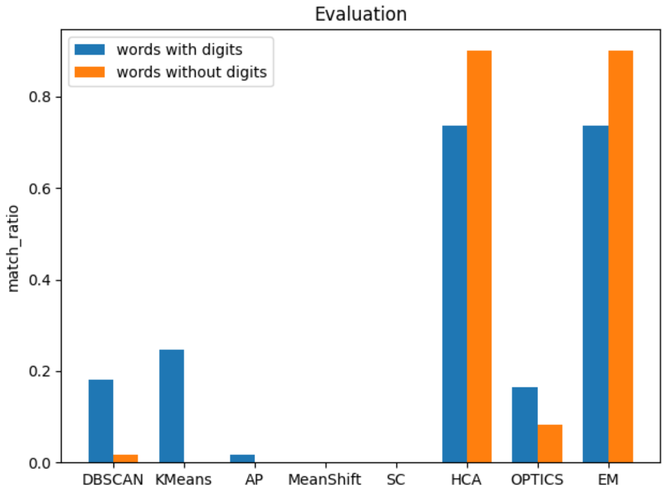

Fig. 2. Evaluation on the clustering algorithms

TABLE I
Comparision between Clustering algorithms

		match ratio	matched words	words in cluster
DBSCAN	without-digits	0.1803	11	461
	with-digits	0.0164	1	10
KMeans	without-digits	0.2459	15	461
	with-digits	0	0	1
AP	without-digits	0.0164	1	3
	with-digits[a]	0	0	0
MeanShift	without-digits	0	0	1
	with-digits	0	0	12
SC	without-digits	0	0	2
	with-digits	0	10	1
HCA	without-digits	0.7377	45	602
	with-digits	0.9016	55	4163
OPTICS	without-digits	0.1639	10	221
	with-digits	0.082	5	799
EM	without-digits	0.7377	45	602
	with-digits	0.9016	55	4163

[a] AP algorithm can't converge the matrix to complete the calculation.

B. *Evaluation*

As mentioned in Section III, we use tfidf method to tranforms the Syslog into tfidf matrix and use the famous clustering algorithms to figure out which combination may be suitable for the automatical Syslog pattern extraction problem for BRAS device. We choose eight famous clustering algorithms [5] and splits the Syslog into two types of dataset and check out whether the words contained digits take the same result as the words without digits produced or not. If those two types of the dataset have the same result, we can imply that our assumption would not correct for this problem-related BRAS device.

Figure 2 shows a result of an experiment for comparing the clustering algorithms with tfidf method. MeansShift and SC algorithms can't extract proper words, both of which produced zero match ratio no matter what type of dataset. It indicates that those combinations would not suitable for this problem.

AP and KMeans algorithms can't extract any word on the dataset that contains words with digits but can extract the words that match the suspicious word list on the other dataset. It implies that those algorithms can't get a significant effect while the Syslog contains digits words and may be suitable for the BRAS extraction problem.

DBSCAN and OPTICS can extract words matched suspicious list but its match ratios are lower than EM or HCA algorithms. We can infer that they use the strict analytic manner to search out a lot of outline data that have different behavior compared with the other words. However, it may not be useful for BRAS Syslog. Because a lot of Syslog words describe the status of device elements or network packets and homogeneity of those words may be not significant compared with the other type of log data. It would cause the algorithms to analyze a lot of clusters.

HCA and EM have higher match ratio values compared with other algorithms. Furthermore, according to the table I, these two algorithms labeled a lot of words to grow up the match ratio. It seems like that they face the same difficulty that DBSCAN and OPTICS faced but they just produce more clusters. By those figures, we may not infer that EM or HCA algorithms are the best clustering algorithm to deal with the problem of automatic extraction BRAS Syslog because these two algorithms produce more words and it could make administrators confused and can't use those words effectively.

V. Conclusion

In this work, we try to combine tfidf method and clustering algorithm to answer the question about how to extract the BRAS Syslog pattern automatically. We conduct an experiment in Section IV-B and the figures show that EM and HCA algorithms have better performance on BRAS Syslog data compared with the other algorithms.

However, those two algorithms can not be proved effective on BRAS Syslog because they try to produce a lot of words to match more words in the suspicious word list that is provided by administrators for testing the clustering combination performances. Moreover, the administrators would feel confused about the huge word list produced by those algorithms because they have to pay more attention to arrange those words and try to pick out the useful words.

Nevertheless, we suggest that EM or HCA algorithms would be better one while we have to deal with the BRAS Syslog pattern extraction problem.

But an issue that we do not consider in this work is that we did not involve data mining methods such as association rule learning [6]. For future work, we can involve data mining methods or consider establish a rigorous suspicious words list to check that which combination can be proved to be a better method to extract the BRAS Syslog pattern.

Acknowledgment

On behalf of all colleagues involved in our team, we are very grateful to those who provided assistance and support to make this work successful.

References

[1] S. E. Hansen and E. T. Atkins, "Automated system monitoring and notification with swatch." in *LISA*, vol. 93, 1993, pp. 145–152.

[2] R. Vaarandi and M. Pihelgas, "Logcluster-a data clustering and pattern mining algorithm for event logs," in *2015 11th International conference on network and service management (CNSM)*. IEEE, 2015, pp. 1–7.

[3] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[4] M. Sanderson, D. Christopher, H. Manning *et al.*, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, p. 100, 2010.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.