Proc. of the 9th Intl. Conf. of Asian Society for Precision Engg. and Nanotechnology (ASPEN 2022)
15–18 November 2022, Singapore. Edited by Nai Mui Ling Sharon and A. Senthil Kumar
doi://10.3850/978-981-18-6021-8_OR-08-0264

Text Mining Framework for Predictive Maintenance in Manufacturing

Duc-Minh Pham[1,#], Su Myat Phyoe[1]

1 Manufacturing Execution and Control, Singapore Institute of Manufacturing Technology (SIMTech), Singapore
Corresponding Author / Email: pham_duc_minh@simtech.a-star.edu.sg

KEYWORDS: Text Mining, Predictive Maintenance, Log files, Maintenance records

In this paper, we propose a framework using integration of text mining algorithms for predictive maintenance in the cyber-physical system of manufacturing. By mining the text of log messages in the monitoring window, information about the coming failure can be predicted. The unstructured log messages are pre-processed and clustered into structured data that makes features can be extracted more efficiently. At the pre-processing stage, data is cleaned up by excluding unnecessary characters and words. A stop words list of system log files is built to remove unnecessary words in log files. The text mining can structure data from different systems version with different log files configuration. Beside features of text content in log messages, other statistical features of log messages are also extracted. The monitoring window is 24-hours, and prediction window time (warning time) is one hour. Maintenance records data is used to extract the failure data to label training data. Timestamp in message log files and maintenance records are used to match these two sets of data. Therefore, with labeling data supervised classification algorithms can be used to predict the failure output. The proposed prediction algorithms not only prediction failure in warning time but also give recommendation for proactive action to prevent failure or preparation before failure happening. The recommendation from data insight can be extracted from text mining in monitoring time and matching with text mining of maintenance records.

1. Introduction

In general, Predictive Maintenance (PdM) can help to decide carrying out of maintenance activities before the failure of equipment occurs. As a result, PdM prevents the failure and helps to minimize breakdown costs and downtime and increase product quality [1]. Description of maintenance data is called maintenance records, or maintenance report, or work orders. Those maintenance records (mainly in text format) normally follow a set of template or guidance, containing a number of attributes such as record id number, asset id number, components, maintenance descriptions (or maintenance activity), cause of failure (if available). There are different types of data and approached to PdM [2-9]. However, in most of the cases, PdM applications use output data collected from sensors that record or monitor physical condition such as temperature or vibration which can be directly linked to the degradation process of the machine. Therefore, sensor data (vibration, temperature, noise etc.) have been well studied and applied in PdM with many publications. However, in some real applications, outputs from sensors are not available and sensors are difficult or costly to be deployed, and in this case event logs generated by the machine are used instead [2-5]. Event-based PdM approaches train models with logged events to recognize patterns which precede a failure incident.

As mentioned, existing state-of-the art use a number of sensor data (e.g. vibration, pressure, current) for PdM which require additional costs to install/maintain the sensors and data labelling effort. Beside sensor data, another source data is text or message data that is generated in even log files or in maintenance records. However, PdM using text mining data has only been mentioned recently in some research due to lack of operation and historical in text. No solution available for fault diagnosis & prognosis for predictive maintenance with log messages or SCADA data which are already available in production line. In [6][7], text mining is used but only applied to maintenance records data. Text mining results can give statistics or occurrence of types (unique “words”) in the text entry field of maintenance records. This study show that text mining can save time in the analysis of data and more information can be extracted that can support decision making. However, it does not show capabilities of text mining to predict failure of assets that need analysis of equipment operation data.

In this paper, we propose a framework using integration of text mining algorithms for predictive maintenance in the cyber-physical system of manufacturing. We combine advantages of two approaches that are described above: text mining for PdM and event logs for PdM. By mining the text of log messages in the monitoring window, information about the coming failure can be predicted. The

unstructured log messages are pre-processed and clustered into structured data that makes features can be extracted more efficiently. At the pre-processing stage, data is cleaned up by excluding unnecessary characters and words. A stop words list of system log files is built to remove unnecessary words in log files. The text mining can structure data from different systems version with different log files configuration. Beside features of text content in log messages, other statistical features of log messages (event logs) are also extracted. The monitoring window is 24-hours, and prediction window time (warning time) is one hour. Maintenance records data is used to extract the failure data to label training data. Timestamp in message log files and maintenance records are used to match these two sets of data. Therefore, with labeling data supervised classification algorithms can be used to predict the failure output. The proposed prediction algorithms not only prediction failure in warning time but also give recommendation for proactive action to prevent failure or preparation before failure happening. The recommendation from data insight can be extracted from text mining in monitoring time and matching with text mining of maintenance records.

2. Architecture of Text Mining Framework for Predictive Maintenance in Manufacturing

There are two data sets operation data and maintenance records data that are used in the proposed PdM framework. The operation data are log files (or SCADA files in some equipment) that contain some attributes:

- Transaction time: when the log file is transacted.
- Transaction identifier: log file identifier that is unique and can be used to search or tag the transaction.
- Event or alarm: this attribute to discriminate the log file whether it is an event or an alarm. Event is normal log file with status or condition of the equipment. Alarm is log file can give a warning about the condition of equipment. Even alarm can give more information about coming failure but both features of even and alarm are studied and extracted to build the machine learning model for PdM.
- Message text: this is important attribute that contains the information text about the condition and status of equipment. Message text can be configured by the equipment manufacturers and may be different from different batch of equipment. This is challenging to standardize or extract the features from text message since it is unstructured data. In the following section, we will propose text clustering and text preprocessing to analyze this message text.
- Alarm code: specify what type of alarm or event code is. Since the alarm code is different number of different batches of equipment. It is challenging if we use alarm code as an input feature for the machine learning model. We propose to use text clustering to re-group the message text or re-group this alarm code into optimal number.

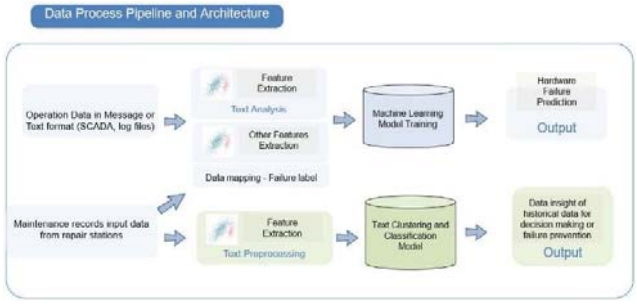

Fig. 1 Architecture of Text Mining Framework for Predictive Maintenance in Manufacturing

Log message data (message text attributes) have noises and are unstructured as followings:

- Two sentences or two phases are connected
- Empty spaces between two words
- Redundant words without valid information
- Configuration of log message in different batch of machines are different

We propose to use text processing and text clustering that are presented in detail in Section 3.

Maintenance records data is another source of data that is used to predict the failure of equipment. There are several important attributes in maintenance records:

- Timestamp: the date and time when failure happens
- Downtime: how long the equipment is down and does not operate normally.
- Category: what type of issues of the failure
- Error: description report of the failure
- Solution: what action of repairing the equipment

Since maintenance records data are written by technicians who serviced the equipment, the data is unstructured with some noise. In our proposed framework as shown in Fig. 1, we find the correlation and matching between operation and maintenance records data:

- Extract the maintenance time (downtime duration) from Error in maintenance records data to find correlation between maintenance and machine alarm/error in operation data.
- Merge (or match) alarm operation data and maintenance records data based on timestamp
- Label the alarms/errors in operation based on Category of maintenance records data

There are two main outputs:

- Predict whether there is failure in a warning window time
- Data insights of historical and operation data in monitoring window time for decision making.

3. Data Pre-processing, Feature Extraction and Machine Learning Models

The unstructured log messages are pre-processed and clustered into structured data that makes features can be extracted more

efficiently. Text pre-processing is implemented as followings:

- Apply regular expression (to split some contents into different columns that are stored in one column: line number (of product line), alarm types, message text)
- Split the connecting words
- Remove punctuation
- Remove digits
- Change to lower case
- Remove stop words
- Stemming (change all the words to root/base word)
- Lemmatization (check with a pre-defined dictionary not to lose the meaning).

Example of stemming and lemmatization are shown in Fig. 2 and the output of text pre-processing is shown in Fig. 3.

Example:

Original Word	After Stemming	After Lemmatization
goose	goos	goose
geese	gees	goose

Fig. 2 Example of stemming and lemmatization

Message_text	clean_msg
Film feeding track 4 film end! STATE: 1	['film', 'feed', 'track', 'film', 'end', 'state']
Film feeding track 4 film end! STATE: 0	['film', 'feed', 'track', 'film', 'end', 'state']
Push in waste box correctly at film loader! STATE: 1	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']
Push in waste box correctly at film loader! STATE: 0	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']
Push in waste box correctly at film loader! STATE: 1	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']
Push in waste box correctly at film loader! STATE: 0	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']
Push in waste box correctly at film loader! STATE: 1	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']
Push in waste box correctly at film loader! STATE: 0	['push', 'wast', 'box', 'correctli', 'film', 'loader', 'state']

Fig. 3 Example of Message Text before and after processed

After text pre-processing, we calculate optimal K-clustering using Elbow Method [10] as described followings:

- For each K, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters K.
- The location of a bend (knee) in the plot as shown in Fig. 4 is generally considered as an indicator of the appropriate number of clusters.

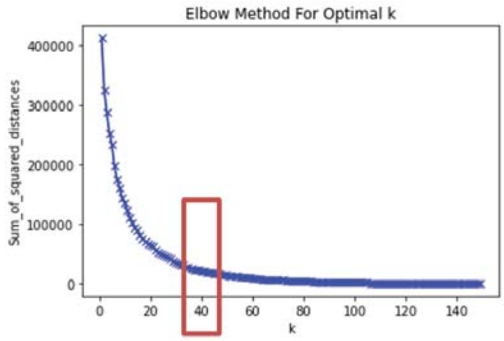

Fig. 4 Optimal Number of Clusters using Elbow Method

After text pre-preprocessing and text clustering, the features from message text and alarm/event can be extracted:

- Statistical features of events
- Statistical features of alarms
- Features of message logs after text preprocessing and text clustering (what types of messages after clustering, length of message logs, numbers of stop words etc.)

One challenge in real operation of maintenance in manufacturing is how to react or prepare when there is an alert of coming failure. Therefore, we propose to use text mining to search and mine information from message text in monitoring window. The text mining can give some recommendations or hints for users to have some proactive action to prevent failure as illustrated in Fig. 5. It also helps them to prepare back up plan to run another production line. Sensor's data may give signals or pattern before the failure could happen, but failure prevention or recommendation is challenging. However, with message text data, the recommendation from data insight can be extracted from text mining in monitoring time and matching with text mining of maintenance records. The proposed approach for mining data in Monitoring Windows for Decision Making is shown in Fig 6. The occurrence of key words or token types can provide information about failure causes, types of failures that may happen in the warning time window. Therefore, these extracted keywords can then be compared and linked to analysis of the other data fields, for additional information to help decision making for prevention of failure.

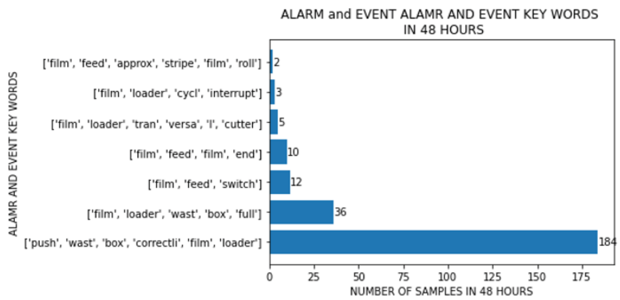

Fig. 5 Example of alarm and event key words from text mining in monitoring window time.

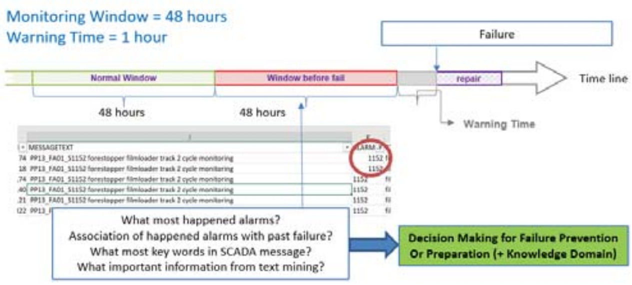

Fig. 6 Proposed approach for mining data in Monitoring Windows for Decision Making

Feature are extracted and used to train the model:

- For each Alarm, calculate 'max, mean, num, std, sum, Skewness , Kurtosis , range, ratio, num of group' of its durations.
- For each group, calculate 'max, mean, num, range, ratio' of the durations. (has 46 alarm groups, consider each group as one)
- For each Alarm, calculate 'max, mean, num, range, ratio' of its intervals.
- For each group, calculate 'max, mean, num' of the intervals.

(has 46 alarm groups from K-clustering, consider each group as one)

There are total 380 features extracted in monitoring window (24-hours log messages data. The data set is 15 months including log messages data (in this case study is SCADA data) and maintenance records data. We use first 12 months of data to train the model and the remaining 3 months to test the model. Machine learning models used for training models are Random Forest and Adaboost. The output of classification model is whether the equipment is failure in the next warning time window 1-hour.

3. Results

Normally, in machine learning classification models, the term "accuracy" is the important performance metric since it has ability to correctly classify all observations in overall.

Accuracy = (True Positive+True Negative)/No of all observation

However, using accuracy to measure performance imbalanced classification can be misleading since, in order to attain high overall accuracy, classifiers would be biased towards the majority class (in this case many output values are "normal"). Therefore, in this PdM application, recall is another good metric to evaluate the classification model's performance.

Recall = True Positive/(True Positive + False Negative)

The result of proposed text mining PdM framework is shown in Fig. 7. The equipment is production line for packaging a product. We ran the data in several lines called LT lines.

Product line station	LT-1	LT-2	LT-3	LT-4	LT-5	LT-6
Train Fail	751	586	818	679	936	1021
Train Normal	4282	3718	4061	3914	3936	3623
Test Fail	493	85	480	360	445	527
Predict Fail Correct	59	0	22	0	17	238
Predict Fail Wrong	434	85	458	360	428	289
Test Normal	1811	488	1841	1771	1621	1444
Predict Normal Correct	1628	488	1815	1763	1445	1091
Predict Normal Wrong	183	0	26	8	176	353
Accuracy	73.22%	85.17%	79.15%	82.73%	76.28%	78.50%
Recall	62.35%	78.50%	70.67%	79.44%	70.76%	67.43%

Fig. 7 Text Mining Framework for Predictive Maintenance in Manufacturing result.

4. Conclusions

In conclusion, a framework using integration of text mining algorithms for predictive maintenance in the cyber-physical system of manufacturing is presented. By mining the text of log messages in the monitoring window and matching with maintenance records using timestamp in two data sets, information about the coming failure in warning time window can be predicted. The unstructured log messages are pre-processed and clustered into structured data that makes features can be extracted more efficiently. Maintenance records data is used to extract the failure data to label training data. The proposed prediction algorithms not only prediction failure in warning time but also give recommendation for proactive action to prevent failure or preparation before failure happening. The recommendation from data insight can be extracted from text mining in monitoring time and matching with text mining of maintenance records.

REFERENCES

1. Ran Y, Zhou X, Lin P, Wen Y, Deng R. A survey of predictive maintenance: Systems, purposes and approaches. arXiv preprint arXiv:1912.07383. 2019 Dec 12.
2. Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014, August). Log-based predictive maintenance. In Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1867-1876).
3. Gutschi, C., Furian, N., Suschnigg, J., Neubacher, D., & Voessner, S. (2019). Log-based predictive maintenance in discrete parts manufacturing. Procedia CIRP, 79, 528-533.
4. Calabrese, M., Cimmino, M., Manfrin, M., Fiume, F., Kapetis, D., Mengoni, M., ... & Toscano, G. (2019, August). An event based machine learning framework for predictive maintenance in industry 4.0. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 59292, p. V009T12A037). American Society of Mechanical Engineers.
5. Calabrese, M., Cimmino, M., Fiume, F., Manfrin, M., Romeo, L., Ceccacci, S., ... & Kapetis, D. (2020). SOPHIA: An event-based IoT and machine learning architecture for predictive maintenance in industry 4.0. Information, 11(4), 202.
6. Stenström, C.; Aljumaili, M.; Parida, A. Natural Language Processing of Maintenance Records Data. Int. J. COMADEM 2015, 18, 33–37.
7. Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. Manufacturing Letters, 27, 42-46.
8. Nota, G., Postiglione, A., & Carvello, R. (2022). Text mining techniques for the management of predictive maintenance. Procedia Computer Science, 200, 778-792.
9. Nota, G., & Postiglione, A. (2021). Text Mining for Industrial Machine Predictive Maintenance with Multiple Data Sources. In Advances in Dynamical Systems Theory, Models, Algorithms and Applications. IntechOpen.
10. Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." Journal of Statistical Software 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.