A Method for Quickly Searching Similar Waveform Patterns in Historical Process Data

Tomohiro Kuroda [*1] *Tetsuya Ohtani* [*2] *Hidehiko Wada* [*3]

Plant operators often rely on their experience when operating plants. However, they could make more appropriate decisions by quickly referring to plant data similar to the current condition among a huge archive of process data. In such cases, efficient referencing of data is crucial.

This paper introduces a technique to satisfy this need by focusing on the waveform patterns in current data and a certain period of historical data. This method can quickly identify and retrieve waveform patterns that match the target ones.

INTRODUCTION

In plants for oil, chemicals, iron and steel, food, and pharmaceuticals, operators often rely on their experience when operating plants, and they can operate plants more efficiently by referring to plant data similar to the current condition among a huge archive of process data. For example, when production is switched from one brand to another with different properties in a continuous production process, the operation data on similar switchovers will serve as a useful reference. In addition, comparing the current process data with abnormalities in the past will help detect abnormalities in the current plant conditions and prevent accidents. To achieve such scenarios, there is a need for technology that quickly identifies conditions similar to the current plant condition among a huge amount of stored operation data.

Time-series process data (trend data) in a specific period can be treated as waveforms. To find similar conditions, it is useful to measure the degree of similarity between the waveforms of the trend data in a target period and the waveforms of the stored trend data.

The degree of similarity between waveforms can be quantified by the Euclidean distance between them(1). However, this is not enough. Even with similar waveforms, when the respective average values of two data sets are different or their phases are shifted from each other, they have a large Euclidean distance and thus are judged to be dissimilar.

Even if a plant keeps producing the same product, there are no trend data sets with identical waveforms. This is because the average of measurement values may vary from season to season, or the time required for chemical reactions changes depending on the amount of materials in tanks, resulting in variations in timing and time-constants of process value change. Therefore, a certain level of tolerance in terms of the measurement value and time must be allowed to quantify the degree of similarity between waveforms.

The dynamic time warping (DTW)(2) algorithm can compare two waveforms having different numbers of data points, allowing shifts in phases and expansion/contraction along the time axis(3). However, DTW sometimes underestimates the distance between not-so-similar waveforms because this algorithm expands or contracts part of trend waveforms arbitrarily (Figure 1). In Figure 1, the horizontal axis is the time axis and the vertical axis represents sensor measurement values. The dotted-line waveform in the figure stays low for a longer time before rising than the solid-line waveform. However, the ascending and descending of the two waveforms look similar and DTW underestimates the distance between the two waveforms. Thus, DTW is not the best way to judge whether two process behaviors are similar or not. A better standard for waveform similarity is needed to identify the essential behavior of processes.

[*1] Incubation Department, Innovation Center, Marketing Headquarters

[*2] RTO Department, Advanced Solutions Center, Advanced Solutions Business Division, Solutions Service Business Headquarters

[*3] Technology Strategy Department, Business Development Center, Marketing Headquarters

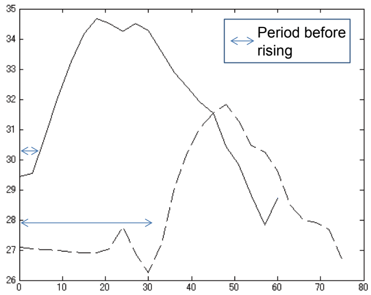

Figure 1 Example of similar waveforms identified by DTW

This paper describes a method of calculating the degree of similarity between trend data waveforms, which is effective for identifying plant conditions. The paper also describes a technology of a quick search to identify waveforms similar to the target waveform (reference waveform) among stored data.

METHOD FOR SEARCHING SIMILAR WAVEFORMS

This section describes a method for searching waveforms that behave similarly, or have a high degree of similarity, to the reference waveform among stored trend data. A method for calculating the degree of similarity, and then a quick search method are described in the following sections.

Method for Calculating the Degree of Similarity

To search for a waveform similar to the reference waveform, we first defined the distance and similarity between waveforms.

Distance between waveforms

In this paper, the Euclidean distance is used for the distance between waveforms. In an actual process, measurement values vary from season to season and the response time and the time until the response starts may change depending on the amount of materials in tanks. When calculating the distance between a certain waveform and the reference waveform, measurement values are offset and time periods are shifted or expanded/contracted if necessary. The resulting minimum distance is set as the distance between the two waveforms. Details of the offset, shift, and expansion/contraction are described below.

- Offset: allows for the difference between measurement values. Specifically, the difference is offset to equalize the average values of the reference waveform and the waveforms in the trend data.
- Shift: allows for the difference between the timing of changes in the time direction. Specifically, when the distance from the reference waveform is calculated, the minimum value of the distance is decided based on the calculation including waveforms with a starting point before and after the original starting point.
- Expansion/contraction: allows for the changes of target period in the time direction. Specifically, when the distance from the reference waveform is calculated, the waveform of a target point is expanded or contracted in the time direction while matching its starting point and ending point to those of the reference waveform. The minimum value is determined as the distance.

Degree of similarity

When the distance between two waveforms is zero, they are judged to be identical. If the distance between two waveforms is large, the waveforms are judged to be different. However, there is no clear standard on how small the distance should be for two waveforms to be judged as similar. To make a clear judgement standard, the calculated distances between waveforms are normalized and the degree of similarity is set between 0 and 100. The maximum degree of similarity (100) means that the two waveforms are identical, and a smaller degree of similarity means that the waveforms are less similar.

Method for Quick Search

If all waveforms among the trend data are compared with the reference waveform in terms of the degree of similarity, every waveform with a high similarity can be searched. However, this calculation takes a long time when the amount of data is large. To solve this problem, we propose a quick search method. In this method, a list of waveforms (index) is prepared from the trend data in advance; these are waveforms representing typical changes occurring in the process (waveform pattern). The searching time is reduced drastically because similar waveforms can be identified only by comparing the reference waveform with the waveforms in the index in terms of the degree of similarity. The procedure of this method is outlined below.

Outline of the algorithm

The procedure of the quick search method is described here. Details of steps (2) through (4) are described later.

(1) Preprocessing: Removing outliers from the trend data, smoothing and compressing are performed.

- Removing outliers: When one value deviates too much from the overall average or from the previous one, the value is set equal to the previous one.
- Smoothing: An exponentially weighted moving average is calculated and applied.
- Compression: An average of several data points is used to represent these values. For example, 60 data sets at one-second intervals are averaged and used as one data set at one-minute intervals.

(2) Preparation of index

(3) Identification of waveform patterns similar to the reference waveform

(4) Comparison of the degree of similarity between the reference waveform and the waveforms in the index

(5) Sorting: The waveforms are sorted in the order of the calculated degree of similarity.

Preparation of the index

The degrees of similarity of all trend data are calculated against the waveform patterns. Among the trend data, waveforms similar to each waveform pattern are registered in advance. In this way, *m* waveforms with a higher degree of similarity are registered in the index for each waveform pattern. Figure 2 shows examples of waveform patterns. Some keep rising consistently, and others rise abruptly. Such waveform patterns are determined in advance.

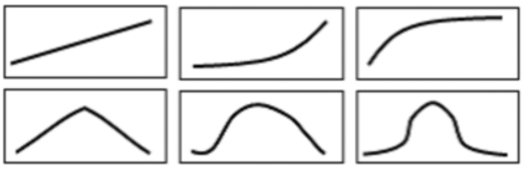

Figure 2 Examples of waveform patterns

Determination of waveform patterns similar to the reference waveform

The degree of similarity of each waveform pattern is calculated against the reference waveform to determine which index should be investigated in detail. *n* waveform patterns with a higher degree of similarity are determined to be similar patterns.

Calculation of the degree of similarity against waveforms in the index

The degree of similarity of all waveforms registered in the index of similar patterns is calculated against the reference waveform. The time required for the calculation is shortened by a factor of *N/nm* compared with calculating the degree of similarity of all waveforms in the trend data. Note that *n, m,* and *N* are the numbers of similar patterns, waveforms registered in the index, and all data points of the trend data after the preprocessing, respectively.

EVALUATION

We evaluated whether the quick search method identified similar waveforms among process data and its accuracy and time required.

Data and Environment for Evaluation

As a process data set for evaluation, temperature data was taken by carrying a thermometer for 47 days. The number of data points, originally approximately 400,000 at 10-second sampling intervals, was reduced to approximately 20,000 by preprocessing. Figure 3 shows part of the data and Table 1 shows the specifications of the PC used for the evaluation.

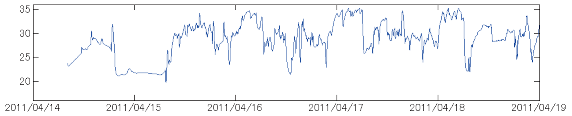

Figure 3 Part of the trend data used for the evaluation

Table 1 Specifications of the PC

CPU	Intel Core i3 (3.07 GHz)
Memory	4 GB
Software	MATLAB R2010b

Evaluation Method

Searching similar waveforms

This evaluation checks whether our proposed method for searching similar waveforms can search waveforms whose behaviors are similar to those of the reference waveform among all data after applying offset, shift, and expansion/contraction preprocessing.

Accuracy of the quick search

This evaluation compares the result of the quick search based on the index and that of the complete search based on the trend data to find whether both searches give similar results. 100 reference waveforms were selected from segments with changes in measurement values. For each reference waveform, the top three search results from the quick search and complete search were compared in terms of the degree of similarity. When the results were identical with each other, the concordance ratio was 100%, and when the results were completely different from each other, the ratio was 0%. Note that the evaluation was based on the combination, not the order, of the top three results. We confirmed that a concordance ratio of 80% or higher is acceptable for immediately searching similar behaviors.

Searching time

The searching times of the two searching methods were also compared by using the same evaluation method described above. We confirmed that a searching time of less than 1 second is acceptable for operators to use during operation.

Results and Discussion

Searching of similar waveforms

Figure 4 shows the result of the complete search. The left side of the figure shows the result with a raw measurement value while the right side shows a measurement value with the minimum distance after the adjustment with offset, shift, and expansion/contraction. The broken and solid lines are reference waveforms and searching results, respectively. Since the raw measurement value has a different length in waveforms (left side of Figure 4), the Euclidean distance cannot be calculated. However, with our proposed method, the distance can be calculated after adjusting waveforms with expansion/contraction. Figure 4 shows that our proposed method can search waveforms whose process behaviors are similar to those of the reference waveform.

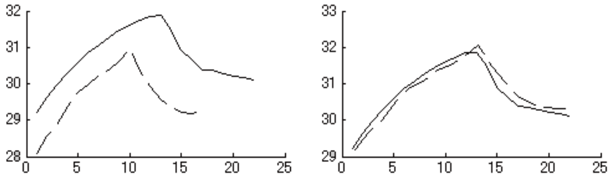

Figure 4 Similar waveforms searched by our proposed method

Accuracy of quick search

The concordance ratio of the top three results from the complete search and quick search was about 70%. Therefore, the quick search can be used to find waveforms with similar behaviors although the configuration of the index needs to be improved.

Searching time

The searching time of the complete search and the quick search was approximately 2.0 second and approximately 0.1 second, respectively. This means that the searching time of the quick search was 20 times faster than that of the complete search. In theory, the searching time should have been improved by 67 times because m = 100, n = 3, and N = 20000 in this evaluation and thus $N/nm \approx 67$. The reason for the difference from the theoretical value is considered to be the overhead time caused by various processes including calling of functions. Although the complete search takes longer than 1.0 second, it is still useful for data analysis.

APPLICATIONS

The method for searching similar waveforms described above was applied to other areas than continuous processes. Examples are introduced below.

Application to Batch Process Data

In analyzing batch process data, it is necessary to extract the data of a specific batch from a massive amount of trend data. This is an easy task if there are flags or other signals to indicate changeovers. Otherwise, the data must be extracted manually, making this task time-consuming. The analysis of batch process data will become more efficient if our proposed method can extract batch data whose waveforms are similar to those of the specified batch data.

We tested this searching method on an actual set of batch process data to see whether it can extract batch data of specific types. We registered the waveform of a batch as the reference waveform, and tried to search waveforms similar to the reference waveform among the trend data. This searching method successfully extracted the data from the same type of batches, including those with slightly shorter or longer reaction times than the reference. This result shows that our searching method can be used for analyzing batch process data.

Application to Other Areas than Process Data

To perform condition-based maintenance on infrastructure such as plants, railroads, roads, and bridges, it is necessary to diagnose their conditions. Whether the condition of facilities is normal or abnormal can be diagnosed by registering waveforms of signals such as current under normal or abnormal operating conditions as the reference waveform in advance, and then evaluating whether the waveforms from real-time measurements are similar to the reference waveform or not.

We conducted a test to see whether the normal or abnormal operating condition of facilities can be judged by comparing the actual operation data with the reference waveforms under normal operating conditions. As seen in Figure 5, waveforms under normal operating conditions show a high degree of similarity to the reference waveform while waveforms under abnormal operating conditions show a lower degree of similarity. Thus, the normal or abnormal condition of this facility can be diagnosed by using the degree of similarity to waveform patterns from its normal operating conditions.

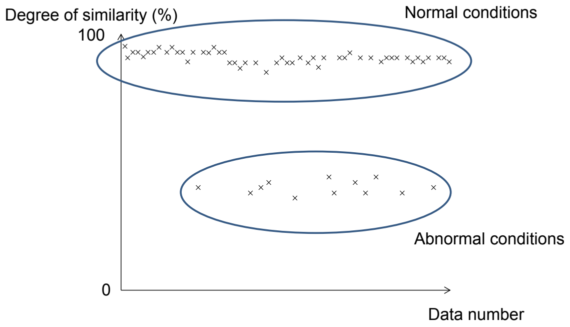

Figure 5 Example of diagnosis

CONCLUSION

We have developed a method to calculate the degree of similarity between waveforms. This method has an excellent sensitivity because it can adjust process data as preprocessing by shifting, extending, or contracting them in terms of measurement values and time. We have also developed a technology to quickly search waveforms similar to the reference waveforms among archived data. The results showed that, in addition to the analysis of process data, our method can be used for other areas such as the diagnosis of facility conditions.

Improvement of the concordance ratio between the result of the quick search and that of the complete search remains a future task. Currently, the index is based on waveform patterns prepared in advance, which include only basic patterns. Therefore, our method may not be able to fully respond to behaviors unique to each plant or changes in process characteristics due to changes in equipment and operation conditions of plants. Thus, waveform patterns should be optimized based on the data from actual plants.

REFERENCES

(1) C. Faloutsos, M. Ranganathan, et al., “Fast Subsequence Matching in Time-Series Databases,” SIGMOD ’94, 1994, pp. 419-429

(2) C. Myers and L. Rabiner, “Level Building Dynamic Time Warping Algorithm for Connected Word Recognition,” IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 29, No. 2, 1981, pp. 284-297

(3) E. Keogh, “Exact indexing of dynamic time warping,” Journal of Knowledge and Information Systems, Vol. 7, Issue 3, 2005, pp. 358-386

* All company names or product names that appear in this paper are either trademarks or registered trademarks of their respective holders.