IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Method of Process Similarity Analysis Based on Longest Sub-similar Sequence Set

To cite this article: Jianhua Zhang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **569** 052036

View the article online for updates and enhancements.

This content was downloaded from IP address 123.138.107.200 on 14/10/2019 at 22:52

IOP Conf. Series: Materials Science and Engineering **569** (2019) 052036 doi:10.1088/1757-899X/569/5/052036

Method of Process Similarity Analysis Based on Longest Sub-similar Sequence Set

Jianhua Zhang, Xiaojun Meng, Huidong Huangfu, Jianglong Zhou, Minjie Xu, Peng Zhang and Jiachen Shen

Northwest Institute of Nuclear Technology, Xi'an, Shaanxi, 710024, China

E-mail: dehuazjh@163.com

Abstract. The traditional process similarity calculation method has a large amount of calculation and the calculation process is cumbersome. For the case of long process routes and long codes, the calculation difficulty and running time increase rapidly, and the operation effect is not good. In this paper, a process similarity analysis algorithm based on the longest sub-similar sequence set is proposed. Compared with the traditional calculation method, this algorithm has the advantages of high operating efficiency and strong adaptability. Further considering the influence of continuous similar process sequences, the penalty factor is introduced to improve the algorithm, so that the algorithm is more perfect and reasonable. This algorithm has significant application value in modern manufacturing systems.

1. Introduction

Process similarity analysis has significant applications in modern manufacturing systems[1-3]. First, group technology is a basic technology of modern manufacturing systems, and the effect of classifying parts into groups is crucial[4]. The effectiveness of grouping depends on whether the judging criteria are reasonable, and the process similarity analysis of parts can provide an objective basis for group technology[5]. Secondly, in the case of new manufacturing requirements, process similarity analysis can help the system to find parts similar to new parts from the existing process library, and provide reference for the preparation of new parts routing[6-8].

In addition, the multi-variety and small-batch production mode has strong flexibility and good market adaptability, and plays a pivotal role in the manufacturing industry[9]. However, in this production mode, the production volume of a single product is small, which makes the production process analysis and control difficult[10]. Using process similarity analysis to integrate and optimize the processing and manufacturing resources of similar parts, a large sample of quality data can be obtained, which is convenient for analysis and control of production process quality.

On the basis of process coding, this paper proposes a process similarity analysis algorithm based on the longest sub-similar sequence set. The algorithm overcomes the cumbersome calculation process and the influence of subjective factors, and eliminates the influence of local process dissimilarity on the overall similarity, and has the advantages of being objective, easy to understand and easy to program. At the same time, this paper considers the influence of continuous similar process sequences, and introduces a penalty factor to improve and perfect the algorithm.

2. Process similarity analysis algorithm based on the longest sub-similar sequence set

2.1. Description of related concepts involved in process similarity analysis

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Assume that the process sequence lengths of two different parts are n_A and n_B respectively. After encoding the process by the coding system, the process code sequences of the two parts are respectively expressed as follows.

$$\begin{aligned} p_A &= [a_1, a_2, a_3, \cdots, a_{n_A}] \\ p_B &= [b_1, b_2, b_3, \cdots, b_{n_B}] \end{aligned} \tag{1}$$

Among them, the elements in P_A and P_B are the process codes of parts *A* and *B* respectively, and the number of element subscripts represent the order of the corresponding process in the product routing.

The set of all sub-similar sequence sets of P_A and P_B is $sp=\{sp_1, sp_2, sp_3,\ldots, sp_x\}$, and any sub-similar sequence set $sp_k=\{\beta_{k1}, \beta_{k2},\ldots, \beta_{kl}\}$ $(k=1, 2,\ldots, x)$ satisfies the following conditions[11].

(1) For any element β_{ki} in sp_k, there are certain elements a_u and b_v corresponding to β_{ki} in P_A and P_B respectively, namely $a_u \rightarrow \beta_{ki}$, $b_v \rightarrow \beta_{ki}$ and $a_u = b_v = \beta_{ki}$.

(2) For the elements β_{ki} and β_{kj} in sp_k, there are a_u and a_v in P_A, if $a_u \rightarrow \beta_{ki}$, $a_v \rightarrow \beta_{kj}$ and $u<v$, then $ki<kj$.

(3) Similar to the previous condition, for the elements β_{ki} and β_{kj} in sp_k, there are b_u and b_v in the set P_B, if $b_u \rightarrow \beta_{ki}$, $b_v \rightarrow \beta_{kj}$ and $u<v$, then $ki<kj$.

When performing the similarity analysis, it is necessary to find the longest sub-similar sequence set, that is, find the element with the longest sequence in the set *sp*. In this paper, the cardinality of the set is used to represent the number of elements in the set, and the cardinality of the set is called the length of the set. The length of the longest sub-similar sequence set of P_A and P_B is as follows:

$$L(p_A, p_B) = \max\{|sp_1|, |sp_2|, |sp_3|, \cdots, |sp_x|\} \tag{2}$$

Example 1: The process code sequences of parts *A* and *B* are "311-312-611-653" and "330-311-810-312-653" respectively, then P_A=[311, 312, 611, 653] and P_B=[330, 311, 810, 312, 653]. (This article takes three-digit process codes as examples.)

In Example 1, all sub-similar sequence sets of P_A and P_B are as follows.

{311}, {312}, {653}, {311,312}, {311,653}, {312,653}, {311,312,653}

The set of sub-similar sequence sets is:

$$sp=\{\{311\}, \{312\}, \{653\}, \{311,312\}, \{311,653\}, \{312,653\}, \{311,312,653\}\} \tag{3}$$

It can be seen that the longest sub-similar sequence set of P_A and P_B is {311, 312, 653} and the length is 3.

As shown below, there are row matrices *M* and *N* of length i_1 and i_2, respectively.

$$\begin{aligned} M &= [m_1, m_2, m_3, \cdots, m_{i1}] \\ N &= [n_1, n_2, n_3, \cdots, n_{i2}] \end{aligned} \tag{4}$$

Definition 1: $M^T \otimes N$ is the $i_1 \times i_2$-order matrix *Q*. For the element q_{kl} of the *k*-th row and the *l*-th column in the *Q* matrix, if $m_k=n_l$, then $q_{kl}=1$, otherwise $q_{kl}=0$.

In Example 1, the result of ${p_A}^T \otimes p_B$ is

$${p_A}^T \otimes p_B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

2.2. Algorithm implementation

In this paper, the similarity of the two process routes is calculated based on the longest sub-similar sequence set. The similarity calculation formula is as follows.

$$sim(p_A, p_B) = \frac{1}{n} \cdot L(p_A, p_B), \; n = \frac{n_A + n_B}{2} \tag{6}$$

Where $L(P_A,P_B)$ is the length of the longest sub-similar sequence set of P_A and P_B.

The value of $sim(p_A,p_B)$ ranges from 0 to 1, and the larger $sim(p_A,p_B)$, the higher the similarity of P_A and P_B. If and only if the conditions $n_A=n_B$ and $a_i=b_i$ $(i=1, 2,\ldots, n)$ are both satisfied, then $sim(p_A,p_B)=1$, which means that the process codes, sequences and lengths of the two routings are completely identical. If and only if Condition 1 is met, then $sim(p_A,p_B)=0$, which means there are no similar processes in the two routes.

Condition 1: For any i, j $(i=1, 2,\ldots, n_A; j=1, 2,\ldots, n_B)$, $a_i \neq b_j$ is satisfied.

It can be seen that the key and difficulty in calculating the process similarity is to obtain the length of the longest sub-similar sequence set.

In the case of short codes and short process routes, the longest sub-similar sequence set and its length can be obtained by manual search. However, some products have long process routes, and if we want to improve the accuracy of process similarity judgment, it is necessary to adopt longer codes. At this time, the workload of manually obtaining the longest sub-similar sequence set is very large.

In order to solve the above problems, this paper proposes an algorithm which is easy to program and solve with computer, and can quickly obtain the longest sub-similar sequence set and its length, and then calculate the process similarity conveniently.

Assume that the process code sequences of parts A and B are $p_A=[a_1,a_2,a_3,\cdots,a_{n_A}]$ and $p_B=[b_1,b_2,b_3,\cdots,b_{n_B}]$, respectively. And the $n_A\times n_B$-order matrix Q can be obtained by the formula $p_A^{\,T}\otimes p_B$. For ease of viewing and display, the results of P_A, P_B, and $p_A^{\,T}\otimes p_B$ are listed below in a table (Table 1).

Table 1. Results representation of matrix Q.

Q	b_1	b_2	b_3	…	$b_{n\mathrm{B}}$
a_1	q_{11}	q_{12}	q_{13}	…	$q_{1n\mathrm{B}}$
a_2	q_{21}	q_{22}	q_{23}	…	$q_{2n\mathrm{B}}$
a_3	q_{31}	q_{32}	q_{33}	…	$q_{3n\mathrm{B}}$
…	…	…	…	…	…
$a_{n\mathrm{A}}$	$q_{n\mathrm{A}1}$	$q_{n\mathrm{A}2}$	$q_{n\mathrm{A}3}$	…	$q_{n\mathrm{A}n\mathrm{B}}$

Where q_{ij} represents the value of the i-th row and the j-th column in the matrix Q, and the row and the column are respectively represented by a two-digit numerical value. For example, the position code corresponding to the value of the third row and the fourth column in matrix Q is 0304.

The basic idea of obtaining the longest sub-similar sequence set of the two process routes using Table 1 is as follows.

First, find all the elements with a value of "1" in Table 1.

Secondly, for any q_{ij} with a value of "1", find all q_{mn} whose value is "1", where m and n satisfy $m>i$, $n>j$. And connect q_{ij} with all qualified q_{mn} in arcs.

Finally, find out the path with the most connected elements, and restore the position code of each element to the corresponding process code to get the longest sub-similar sequence set (the longest sub-similar sequence set may exist more than one at the same time).

According to the above method, the result representation of the basic idea of obtaining the longest sub-similar sequence set of Example 1 can be expressed in the form of Table 2.

Table 2. The result representation of the basic idea of obtaining the longest sub-similar sequence set.

Q	**330**	**311**	**810**	**312**	**653**
311	0	1	0	0	0
312	0	0	0	1	0
611	0	0	0	0	0
653	0	0	0	0	1

As can be seen from Table 2, the position codes of the longest sub-similar sequence set of P_A and P_B are [0102, 0204, 0405], and the corresponding process codes are {311, 312, 653}.

The above method transforms the analysis of the process codes into the analysis of the position codes. For the case of the complex process codes, it can be finally converted into a four-digit position code for comparison and analysis, which greatly simplifies the calculation process.

Based on the above ideas, this paper proposes an algorithm for finding the longest sub-similar sequence set and its length. The algorithm steps are as follows.

L=2;Q=zeros(n_A,n_B);K=ones(1,n_B);

Step1. *for (i=1:n_A) {*
for (j=1:n_B) {
if (P_A(i)==P_B(j))
then {
Q(i,j)=1;
*J(K(1),1,1)=i*100+j;*
K(1)=K(1)+1;
}}}

Step2. *While (K(L-1)>1) {*
for (l=1:(K(L-1)-1)) {
for { i=(J(l,L-1,L-1)/100+1):n_A } {
for { j=(J(l,L-1,L-1)%100+1):n_A } {
if (Q(i,j)=1)
then {
*J(K(L),1:L,L)=[J(l,1:(L-1),L-1),i*100+j];*
K(L)=K(L)+1;
}}}}
L=L+1;}

Step3. *L=L-2;sim=2*L/(n_A+n_B);*

In the above algorithm, L represents $L(P_A,P_B)$, and *sim* represents $sim(p_A,p_B)$.

In Example 1, $L(P_A,P_B)=3$ and $sim(p_A,p_B)\approx0.67$, representing that the similarity of process route A and process route B is 0.67.

3. Process similarity analysis algorithm with penalty factor

In Example 1, the longest sub-similar sequence set of P_A and P_B is "311-312-653", and the process similarity $sim(p_A,p_B)\approx0.67$.

Example 2: For two parts A_1 and B_1, where P_{A1}=[330,810,311,312,653] and P_{B1}=[311,312,653,611], the longest sub-similar sequence set of the two is "311-312-653", and process similarity $sim(p_{A1},p_{B1})\approx0.67$.

The longest sub-similar sequence sets of Example 1 and Example 2 are both "311-312-653" and the process similarities are both 0.67. The longest sub-similar sequence set "311-312-653" of Example 2 is a sequence of consecutive similar processes, which should have a greater degree of similarity than Example 1. Therefore, it is not accurate to calculate the similarity of the two process routes by only relying on the longest sub-similar sequence set.

In order to solve the above problems, this paper considers the effects of consecutive similar process sequences, introduces a penalty factor, and proposes an improved algorithm for process similarity analysis.

The improved algorithm formula can be expressed as:

$$sim'(p_A, p_B) = (\alpha + 1) \cdot sim(p_A, p_B) \tag{7}$$

In the above formula, α represents the penalty factor, and

$$\alpha = [1 - sim(p_A, p_B)] \cdot \frac{1}{n} \sum_{i=1}^{k} (R_i - 1) \tag{8}$$

In formula (8), R_i is the length of the i-th consecutive similar process sequence of a certain longest sub-similar sequence set, and k is the number of consecutive similar process sequences. (R_i = 1 indicates that the length of the i-th consecutive similar process sequence is 1, that is, there are no two or more consecutive similar processes.)

3.1. Analysis of rationality of improved algorithm

The rationality of the improved algorithm formula is analyzed in three cases.

(1) When $sim(p_A,p_B)=0$, it is easy to get $sim'(p_A,p_B)=0$, that is, the process similarity obtained by equation (6) is 0, and the similarity calculated by the improved algorithm (equation (7)) is also 0, indicating that there are no similar processes between the two process routes.

(2) When $sim(p_A,p_B)=1$, then $\alpha=0$ and $sim'(p_A,p_B)=1$, that is, for the two process routes that are completely similar, the similarity calculated by the improved algorithm is also 1.

(3) The case of $0<sim(p_A,p_B)<1$ is discussed as follows.

In the case of $0<sim(p_A,p_B)<1$, the result obtained by the improved algorithm should satisfy $sim'(p_A,p_B)\geq sim(p_A,p_B)$ due to the influence of the consecutive similar process sequences. In addition, the maximum value of similarity is 1, so in the case of $0<sim(p_A,p_B)<1$, the improved similarity should also satisfy $sim'(p_A,p_B)<1$. It is easy to prove that both of the above conditions can be satisfied, so for the case of $0<sim(p_A,p_B)<1$, $sim(p_A,p_B)\leq sim'(p_A,p_B)<1$ can be obtained, which is in line with the actual situation, and is also consistent with the original intention of the improved algorithm proposed in this paper (there should be a greater degree of similarity when there are consecutive similar process sequences, that is, the process similarity should be slightly larger than the case without consecutive similar process sequence).

3.2. Improved algorithm implementation

The key to calculating the process similarity using the improved algorithm formula is to obtain all consecutive similar process sequences and their lengths to determine the value of the penalty factor α.

The set of position codes for all of the longest sub-similar sequence sets can be obtained in Section 2.2, as shown below.

$$numsp_{\max} = \{num_1, num_2, \cdots, num_w\} \tag{9}$$

Any element in the above set is a set of position codes corresponding to a certain longest sub-similar sequence set. Without loss of generality, only one of the elements is analyzed below.

For any element $num_r=[c_1, c_2,\ldots, c_L]$ ($L=L(P_A,P_B)$; $r=1, 2,\ldots, w$) in set $numsp_{max}$, any element c_j ($j=1, 2,\ldots, L$) in num_r is a four-digit number. The first two digits of c_j indicate the row in which the j-th similar process is located in the matrix Q, that is, the positional order in $p_A=[a_1,a_2,a_3,\cdots,a_{n_A}]$. The last two digits indicate the column in which the j-th similar process is located in the matrix Q, that is, the positional order in $p_B=[b_1,b_2,b_3,\cdots,b_{n_B}]$.

In Example 1, there is only one longest sub-similar sequence set, namely $numsp_{max}=\{num_1\}$, where num_1=[0102, 0204, 0405].

Definition 2: For c_{j+1} and c_j ($j<L$), if $c_{j+1}=c_j+0101$, then c_{j+1} and cj are said to satisfy the "continuous condition".

The "continuous condition" can be used to determine whether any two adjacent processes in the longest sub-similar sequence set are continuous similar processes. The algorithm flow for determining all consecutive similar process sequences and their lengths is shown in Figure 1.

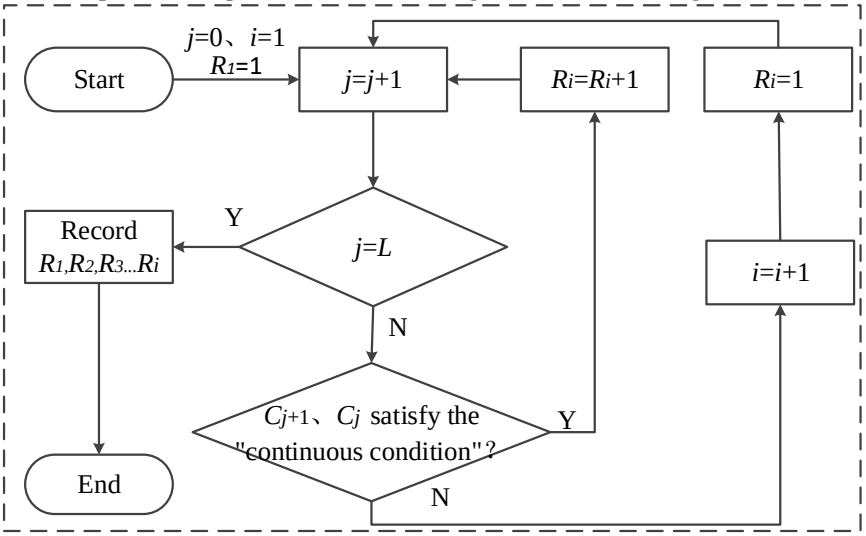

Figure 1. Algorithm flow for calculating *Ri.*

The steps of improved algorithm with penalty factors are as follows.

Step4. *for (t=1:(K(L)-1)) { j=0;i=1;R(i)=1;*
for (j=1:(L-1) {
if (((MtxP(t,j)/100+1)==(MtxP(t,j+1)/100))
&& ((MtxP(t,j)%100+1)==(MtxP(t,j+1)%100)))
then {R(i)=R(i)+1;}
else {i=i+1;R(i)=1;}}
R=R(1:i);
alpha(t)=(2/(nA+nB))(sum(R)-i);*
*sim2(t)=(alpha(t)+1)*sim;}*

In the above algorithm, *sim* represents $sim(p_A,p_B)$, and *sim2(t)* represents the set of process similarities for different longest sub-similar sequence sets.

4. Application examples

Example 3: The process code sequences of the two parts A_2 and B_2 are as follows.

$$p_{A2}=[311,312,313,314,315,452,856,726,734,843,812,624]$$
$$p_{B2}=[311,313,734,843,726,953,313,314,452,812,621,545] \quad (10)$$

The result of Example 3 calculated by the above algorithm is shown in Figure 2.

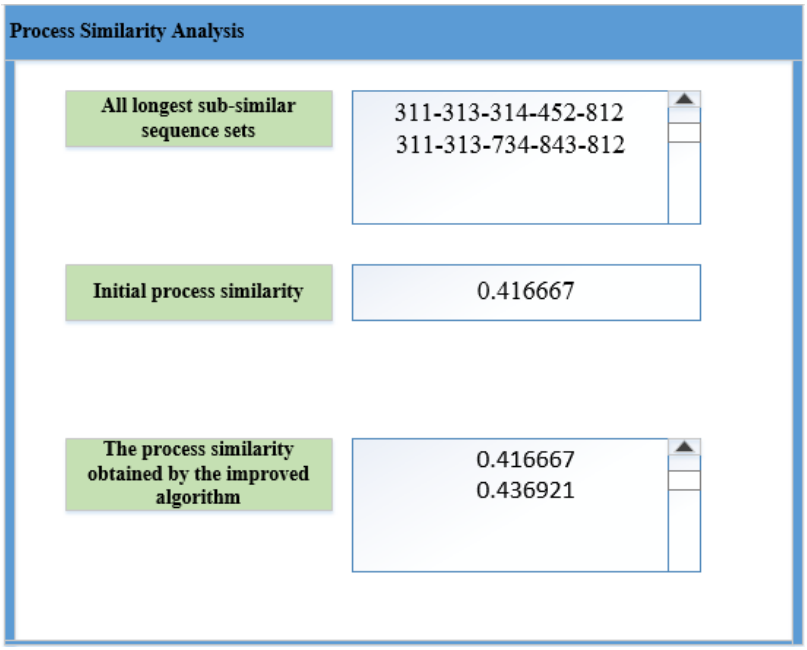

Figure 2. Result of Example 3.

It can be seen from Figure 2 that the length of the longest sub-similar sequence set of P_{A2} and P_{B2} is 5, and there are two longest sub-similar sequence sets, as shown below.

{ 311, 313, 314, 452, 812}
{ 311, 313, 734, 843, 812}

The process similarity calculated by equation (6) is 0.416667, and the process similarities obtained by the improved algorithm (equation (7)) for the two different longest sub-similar sequence sets are 0.416667 and 0.436921, respectively. Since the "734-843" in the second longest sub-similar sequence set is a sequence of continuous similar processes, the similarity calculated from the second longest sub-similar sequence set is slightly larger than the similarity calculated from the first.

5. Conclusion

In this paper, a process similarity analysis algorithm based on the longest sub-similar sequence set is proposed. This algorithm can eliminate the influence of local process dissimilarity on the overall similarity and can overcome the shortcomings of traditional algorithms that require a large amount of data comparison and calculation. The algorithm not only is easy to program, but also can be applied to calculating the process similarity quickly for the case of long codes and long process routes. In addition, considering the high local similarity of the process, the penalty factor is proposed based on the consecutive similar process sequences, and the algorithm is improved to make it more perfect and reasonable. The research results of this paper can be applied to group technology, manufacturing resources optimization and process quality analysis and control of multi-variety and small-batch production, and has important application value in modern manufacturing systems.

References

[1] Song H, Zuo D, Xue S, Jiao G. (2009) Research on rapid design technology of final assembly process based on process similarity. China Manufacturing Informatization, 38(21): 29-31+35.
[2] Wei Q. (2003) Research on the similarity judgment technology of parts in CAPP. Sichuan University.
[3] Liu W, Liu Z, Tan J. (2010) Product module construction method based on process similarity and its application. Journal of Computer-Aided Design & Computer Graphics, 22(10): 1647-1654.

[4] Liu W, Zhou H, Jing X. (2006) Code-based new part grouping process matching in production process analysis method. Modern Manufacturing Engineering, 10: 56-58+113.
[5] Zhang Y. (2003) Comparative analysis of similarity degree of parts classification into groups. Journal of Chengdu University(Natural Science Edition), 02: 44-46.
[6] Zhou Y, Mei Z. (2016) Application of improved similarity algorithm in retrieval of composite component molding process. Aeronautical Manufacturing Technology, 08: 76-80+84.
[7] Xu C, Qian S, Huang F, Yang Y. (2016) A extension correlation retrieval method for process similarity of box parts. Machinery Design & Manufacture, 09: 173-175+179.
[8] CHEN J, GONG Z, LIU L, CHENG Y. (2012) Study on retrieval method of parts based on process similarity. Machine Tool & Hydraulics, 40(01): 42-47.
[9] Chen X, Chen F. (2019) Bayesian dynamic quality control method for multi-variety and small-batch production. Journal of Xi'an Jiaotong University, 06: 1-7.
[10] Zhao H, Zheng C, Zhao W. (2018) Compilation method and quality control measures for multi-variety and small-volume production planning [J]. Chinese and Foreign Entrepreneurs, 09: 195.
[11] Chen Q. (1999) Comparison Method of Similar Strings. Journal of South China Normal University (Natural Science Edition), 02: 37-41.