IOP Conf. Series: Materials Science and Engineering **382** (2018) 052004 doi:10.1088/1757-899X/382/5/052004

Complex network model of process industry based on big data

Huan Jun PANG, Yu Shui GENG*, Xue Song JIANG

Information Institute Qilu University Of Technology, Ji'nan, Shandong 250000, China

Email: UltraTab@outlook.com; gys@qlu.edu.cn; 37159700@qq.com

Abstract: The production process of the process industry is generally a continuous or batch process, with the feature of complexity, uncertainty, multi-objective, multi-constraint, multi-resource coordination and so on. With the advent of industry 4.0 and the era of big data, the process industry began to be developed in the direction of digital automation, at the same time, the idea of big data brings us a new direction to solve the process scheduling problem. It has a huge impact on the results of the optimization that the way of organizing and using these data which comes from production process. In this paper, the industrial process of producing glass fiber by alkali-free kiln process has been chosen as an example, the characteristics and shortcomings of the task - source network model are analyzed and a complex network model of process industry scheduling based on big data has been introduced, which accurately links the data in the production process.

1. The characteristics of the process industry big data

Process industries refer to industries such as petrochemicals, electricity, metallurgy, and so on, which are characterized by continuity. [1]. Process Industrial production scheduling objectives include economic indicators and performance indicators, and ultimately reflected in the lowest cost or the most profitable. The general trend is automation, centralization and integration [2].

Process Industry Big Data is different from other types. First, its data volume is much smaller than other types of big data: Facebook generates as much as 500TB of data per day, while a production process may generate only a few GB of data per day. However, its capacity, rate and type are huge compared to previous industrial data, so we still regard it as big data. At the same time, the production process of the process industry is relatively stable, resulting in a higher degree of data redundancy, mining the valuable data relative to other types of large data more difficult.

2. Mathematical description of complex networks

2.1 Complex network

In this paper, we will describe and evaluate the established model in the following areas.

A network G can be denoted G=(N,E) with node set N and edge set E, where node set $N=\{i|i=1,2,3,\cdots,n\}$.In this paper, G is a directed network, then the edge set can be expressed as $E = \{e_{ij} | i, j (1, 2, 3, ..., n)\}$ where e_{ij} represents the edge from i to j.[5]

The weight of the edge W reflects the intensity of the edge. If the edge in network G is not equivalent, we need to increase the weight set W on this basis. In this paper, G is a directed network, then the edge set can be expressed as $W=\{w_{ij}|i,j\in(1,2,3,\cdots,n)\}$ where e_{ij} represents the weight of the edge where the starting point is i's end point j.

The degree is a simple but important concept of the properties of a single node. The degree k_i of node

i is defined as the number of other nodes connected to the node. The degree of a node in the network is divided into in-degrees and out-degrees. The average of the degree k_i of all nodes i in the network is called the average of the network, denoted as <k>. The distribution of the degree of nodes in the network can be described by the function P (k). P (k) represents the probability that a randomly selected node is exactly k.[6]

The distance d_{ij} between two nodes i and j in the network is defined as the number of edges on the shortest path connecting the two nodes. The average path length L of the network is defined as the average of the distances between any two nodes,

$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \geq j} d_{ij} \quad (1)$$

where N is the number of network nodes.

A node in the network has k_i edge and it and other node necklaces, which are called neighbors of node i.[7] Obviously, there may be $k_i(k_i-1)/2$ edges between the k_i nodes. And the ratio of the number of edges E_i and the total possible number of $k_i(k_i-1)/2$ between the k_i nodes is the clustering coefficient C_i of node i:

$$C_i = 2E_i/k_i(k_i-1) \quad (2)$$

The clustering coefficient C of the whole network is the average of the clustering coefficients Ci of all nodes i:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i \quad (3)$$

2.2 Apriori algorithm

The formal description of the association rule mining problem is: Let $i = \{i_1, i_2, ..., i_m\}$ be a collection of items. Let the task-related data D be the transaction database, where each transaction T is the set of items, making $T \in I$. Each transaction has an identifier TID. A collection of multiple items is called an item set. An item set consisting of k items is called a k item set. Let A be an item set, and transaction T contains A if and only if $A \in T$. The association rule is an implicit expression like $X \rightarrow Y$, where $X \in I$, $Y \in Y$ and $X \cap Y = \emptyset$. The strength of the association rule can be measured with its support and confidence[10]. The support degree determination rule can be used to give the frequency of a given data set, and the confidence determination determines the frequency of occurrence of Y in the transaction containing X. The degree of support (s) and confidence (c) are defined as follows:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (4)$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{Y} \quad (5)$$

Where $\sigma(X) = \{T_i | X \in T_i, T_i \in D\}$, N is the total number of transactions. If $s(X \rightarrow Y) \geqslant$ minsup (minimum support) and $c(X \rightarrow Y) \geqslant$ minconf (minimum confidence), the association rule is a strong association rule, otherwise it is called weak association rule.

3. Task - resource flow network model

In recent years, some scholars have suggested that some problems of complex system objects in the process industry are also suitable for the analysis of the ideas and methods of complex network theory because of their complex association of function and structure. Examples of existing research include network modeling analysis of process workflows such as refinery ammonia, and some research results have been made.

3.1 Establishment of Model

In this paper, the process of producing glass fiber from alkali-free kiln process is taken as an example, based on the analysis of network structure characteristics, this paper focuses on the analysis of the structure - function relationship, and links the topology of the network with the industrial production characteristics of the object better.

In the task-resource flow network model[8] of the glass fiber production industry process, the node represents a task, and the edge represents the transfer of resources between the two tasks. For example, the processing task by dealing with a variety of raw materials, including raw materials or semi-finished products, produce the corresponding product, and pass it to the downstream other tasks.

The logistics of various tasks, energy flow and information transmission are unified as a resource exchange. According to the specific production program, you can quantify the logistics between the task unit exchange, which is defined as the weighted edge of the network. According to the material balance of the system, the input of each task unit is equal to its output.

According to the above modeling rules, the task-resource flow network in this paper is a directed weighting network. The orientation of the edge is determined by the direction of movement of the material. The weight is determined by the production scheme. Each production scheme specifies the distribution ratio of the discharge capacity of each task unit. The same task unit feeds is equal to the discharge quantity.

3.2 Model analysis

The model contains 53 nodes, and the structural parameters of the model are shown in the following table

Table 1 The structural parameters of task - resource flow network model

N	E	L	C	$<K_{in}>$	$<K_{out}>$
53	97	2.8062	0.093053	1.8302	1.8302

The degree of each node and the degree distribution of nodes are shown in the following figure:

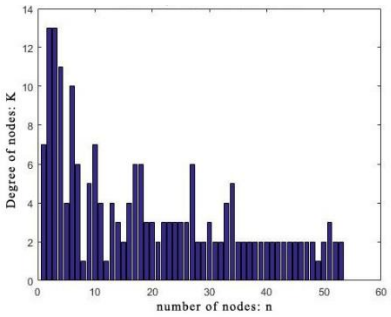

Figure 1 Degree of nodes in Task-resource flow net-work model

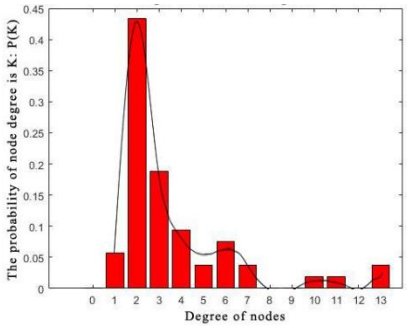

Figure 2 Degree distribution diagram

It can be seen from the figure that the degree distribution of the network shows the phenomenon of "long tail", showing the characteristics of power law distribution, with the nature of scale-free network, most of which nodes are small, Greater degree, but the number is less. The nodes that have bigger degrees are more important than others, and their changes can cause tremendous impact on the network [9]

3.3 Disadvantages of the model

The task-resource network model truly reflects the status of the production units in the whole production process of the glass fiber production process. The important nodes found in the generated simplified network are in line with the important links in the actual production and can objectively react to the production The process is a successful model. However, the model still has some shortcomings, for example：

- It can not reflect the dynamic process of production. The production process in the process industry is not immutable, and once the production plan is changed, the weight in the network will change and the original network would not be able to respond to the new production plan.

- The model can't accurately reflect the use of energy. In the actual production process, in addition to the flow of resources, the use of energy is also a link can't be ignored, only for resource flow optimization results might be the local optimal solution.

4. Complex network model based on data

4.1Conditions of Model establishment

With the advent of the Industrial 4.0 era, the production process is more intelligent in the context of information network and knowledge resource era, and it is also more dependent on various data.[11] With the arrival of big data era, the process industry has entered a large data age, the process of industrial production process generated data per day can reach GB level, compared with the past process industry, the amount of data generated is very large.[12] The large amount of data generated in the production process provides the conditions for the establishment of the model.

4.2 Establishment of Model

4.2.1 Nodes settings

The data-based flexible shop network G=(N,E,W) is established on the basis of the process entity, where node $N=\{n_i|i=1,2,3,\cdots,n\}$ represents the data returned on the sensor of the process node. A process node may contain multiple sets of data, for example, a batch preparation node may contain temperature, wind speed, oxygen concentration and other sets of data.

The edges of the data nodes $E=\{e_{ij}|i,j\in(1,2,3,\cdots,n)\}$ are directed whose direction is based on the production process and the edge can only be directed by the upper node or peer node Or a peer node, which can not point to the parent node by the lower node. The relationship between the data node and the process entity is shown in the following figure:

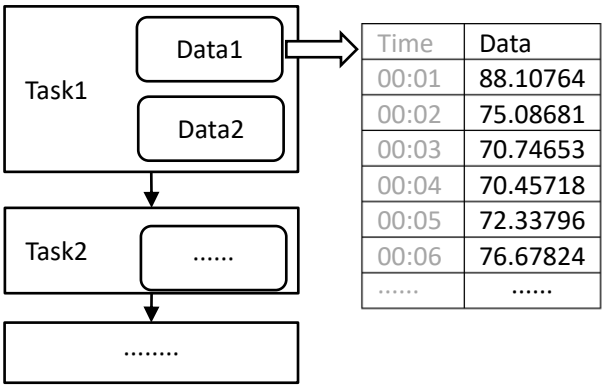

Figure 3 The relationship between the tasks and the data nodes

4.2.2Edges Setting

1) Data preprocessing

We set every group of data generated from each data nodes as a time-ordered vector: Let N_1 and N_2 be the set of data nodes of the adjacent production units in the production process, N_1 contains the elements where the production unit is located upstream of N_2.In the column vector, the data that is increased compared to the previous data is replaced by +1， the data with the same value as the previous data is replaced by 0， and the data with the reduced value compared to the previous data is replaced by -1, this will make each set of column vectors to contain only three events including increasing, decreasing, and remaining unchanged.

2) Association rules mining

Data nodes between two adjacent task nodes build new edges through Apriori algorithm, that is:

We assume that a change in data node data only affects the data set in the set and the adjacent downstream collection, so we only analyze the data nodes contained in the adjacent two production units. Let A, B for the two columns in line with the conditions of the vector, they have the state of the following circumstances the same time:

Table 2 All possible combinations of events

B \ A	1	0	-1
1	P(1,1)	P(0,1)	P(-1,1)
0	P(1,0)	P(0,0)	P(-1,0)
-1	P(1,-1)	P(0,-1)	P(-1,-1)

We further simplify:

Table 3 Possible associations

Result	Detail	All possible circumstances
A,B have association	A→B	P(1,1)∪P(0,0)∪P(-1,-1)
	A→-B	P(1,-1)∪P(0,0)∪P(-1,1)
A,B have no association	Others	P(1,0)∪P(0,1)∪P(0,-1) ∪P(-1,0)

We can easily see that these three events can not happen at the same time,

If any association rule of A → B or A → -B is established, we think that A and B are related. We set the minimum support for 10%, the minimum confidence of 75%, that is:

$$\begin{cases} s(A \to B) = s(1,1) + s(0,0) + s(-1,-1) \\ s(A \to -B) = s(1,-1) + s(0,0) + s(-1,1) \\ s(others) = s(1,0) + s(0,1) + s(0,-1) + s(-1,0) \end{cases} \quad (6)$$

$$\begin{cases} c(A \to B) = c(1,1) + c(0,0) + c(-1,-1) \\ c(A \to -B) = c(1,-1) + c(0,0) + c(-1,1) \\ c(others) = c(1,0) + c(0,1) + c(0,-1) + c(-1,0) \end{cases} \quad (7)$$

If A, B has a strong association rule, then the edge between A and B will be established, otherwise there is no edge exist between A and B.

3) Weight setting

In order to express the data relationship between nodes more accurately, in this model, we allow negative weights of the edges. The weight set of the edge can be expressed as $W=\{w_{ij}=f(n_i,n_j)|i,j\in(1,2,3,\cdots,n)\}$, w_{ij} is obtained by the function f (n_i, n_j), the greater the effect of the edge change on the end point, the greater the absolute value of the edge weight. To simplify the model, we assume that the effects between adjacent nodes are linear, therefore:

$$w_{ij} = \frac{1}{n}[n_i]^T \cdot [n_j]^{-1} \quad (8)$$

In addition, $[n_i]$ is the column vector, the value of n is equal to the number of data in the column vector.

4.2.3 Model analysis

Compared with the task-resource network, the data-based complex network model contains more nodes, so its network size is much larger than the task resource network. At the same time, the model can digitize the entire production process, construct a network composed of numbers, the relationship between the network for the relationship between different data, which is more convenient for the parameters to optimize.

1)Parameters of model structure

The structural characteristics of the complex network model based on data are shown in the following table

Table 4 The structural parameters of data based complex network model

N	E	L	C	<Kin>	<Kout>
124	243	3.0926	0.10327	1.6613	2.2581

As can be seen from the above table, the data-based Complex Network Model for Glass Fiber Fabrication is large than the task - resource network, besides that out-degree and in-degree is not symmetrical either. This is because many of the data in the production process changes are the need for human intervention, and not affected by other nodes, so these nodes only exist out without the entry.

At the same time, the complex network model based on data also shows the characteristics of power law distribution, the degree distribution as shown below:

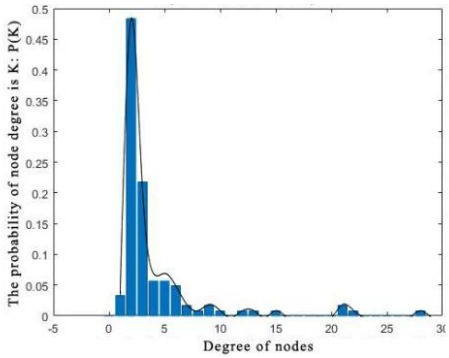

Figure 4 Degree distribution diagram of Task - resource flow network model

2)The advantages of the model

In the task-resource network model, we can accurately identify the important nodes in the production process，witch always have the obvious structural features, for example, the degree of node is relatively large. These nodes play an important role in the production process, so we need to devote more effort to these nodes during production. However, the task-resource network is not flexible enough for multi-objective optimization problems.

In the data-based complex network model, the nodes are composed of data, while the edge of the network reflects the relationship between the data, by the model, we can optimize the process to optimize the data, the main advantages are the following points:

- Shield the complex production processes：The model only reflects the link between the production data and does not focus on which part of the data belongs to, and the optimization of the process is translated into the optimization of the data.
- Maximize production data：The data type of industrial data and its' source is different to the data generated by the Internet or the transportation network. Industrial data is usually pure data, and the data redundancy is high, so when dealing with large industrial data, the data screening should pay special attention. The model eliminates the need for manual filtering of data during modeling, avoiding the loss of important data
- Optimization is more accurate: In the production process, a production link may contain a number of data. In the previous optimization program, we always artificially select the high importance of the parameters and abandon the importance of low parameters with the help of field expert. While reducing the amount of optimization work, but also lost lots of data, making the optimized results and the optimal solution have a certain deviation. In this model, our optimization object is no longer a production unit, but a data, so the optimization is more targeted, and more accurate.

5. Summaries

This article describes a complex network of data nodes that uses a large number of data sets generated in industrial production in the context of Industry 4.0 to build complex networks. Using complex network models based on data, we can transform the optimization of the process into the optimization of the data. At the same time, the model effectively shields the production process, and the optimization staff only needs to analyze the relationship between the data without analyzing the detail of production process. The model also more intuitively expressed the relationship between the various parameters in the production, and provided a new way to solve the problem of process optimization.

Acknowledgement
This work was supported by Shandong Provincial Natural Science Foundation, China (NO. ZR2013FM017)

Reference

[1] Zhao X Q, Rong G. Survey of Production Scheduling in the Process Industry[J]. Control & Instruments in Chemical Industry, 2004.
[2] Xia M S. Research on Intelligent Plant Construction Technology of Process Industry[J]. Information Technology & Informatization, 2013.
[3] König D. Theory of Finite and Infinite Graphs[M]. Birkhäuser Boston, 1990.
[4] Erdos, P. and A. Renyi, On random graphs. Publicationes Mathematicae Debrecen,1959.6：p.290-297
[5] WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Network science: an introduction[M]. Beijing: Higher Education Press, 2012.
[6] Deng W, Li W, Cai X, et al. The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(8):1481-1485.
[7] Yang M, Zhang J Y, Zhang D M. Overview on Network Models of Complex Networks Topology Structure[J]. Communications Technology, 2014.
[8] Pantelides, C.C. Unified Frameworks for optimal process planning and scheduling. In Proceedings on the second conference on foundations of computer aided operations. 1994: Cache Publications New York.
[9] Goh, K.-I. Review E. et al., Betweenness centrality correlation in social networks. Physical 2003 . 67(11:.017101.
[10] Lin C, Yangyang W U, Huang Z, et al. Parallel Research of Apriori Algorithm Based on MapReduce[J]. Journal of Jiangnan University, 2014, 1(2):236-241.
[11] Architexturez. PR: Industrial Convergence Revolution: Smart Cities, Industry 4.0, Big Data in Manufacturing, and Industrial IoT[J].
[12] Wang H, Osen O L, Li G, et al. Big data and industrial Internet of Things for the maritime industry in Northwestern Norway[C]// TENCON 2015 - 2015 IEEE Region 10 Conference. IEEE, 2015:1-5.