World Wide Web (2020) 23:1407–1421
https://doi.org/10.1007/s11280-019-00707-8

Master data management for manufacturing big data: a method of evaluation for data network

Chun Zhao[1,2] · Lei Ren[2] · Ziqiao Zhang[2] · Zihao Meng[2]

Received: 7 February 2019 / Revised: 7 May 2019 / Accepted: 8 July 2019 /
Published online: 16 July 2019

Abstract
In the process of manufacturing, a large amount of manufacturing data is produced by different departments and different domain. In order to realise data sharing and linkage among supply chains, master data management method has been used. Through master data management, the key data can be shared and distributed uniformly. However, since these cross-domain data form a data network through the association of master data, how to evaluate the effectiveness and rationality of this network becomes the major issue in the proposed method. In this paper, a model of the master data network is built based on the theory of set pair analysis. In order to verify the master data, an evaluation method for the network is proposed. Finally, a case was presented to validate this network model and evaluation method.

Keywords Cloud manufacturing · Master data · Set pair analysis · Modeling and simulation

1 Introduction

With the rapid growth of cyber technology, the boundary between cyberspace and the physical world begins to fuzzy. Through the interaction of cyberspace with the physical world, the accurate prediction of behaviour and the rapid development of systems can be implemented [35]. This concept is being applied rapidly in the manufacturing field, Cyberization has become an important topic in manufacturing. Thereby, some concepts related to smart computing and smart manufacturing have been proposed, such as Cloud Manufacturing [12], Industry 4.0 [11], Industrial Internet of Things [7], Cyber-Physical-Social Systems [21] and so on. With the growing popularity of cloud manufacturing concepts [32], more

This article belongs to the Topical Collection: *Special Issue on Smart Computing and Cyber Technology for Cyberization*
Guest Editors: Xiaokang Zhou, Flavia C. Delicato, Kevin Wang, and Runhe Huang

✉ Lei Ren
lei_ren@126.com

[1] School of Computer, Beijing Information Science and Technology University(BISTU), North 4th Ring Mid Street 35, Beijing, China

[2] School of Automation Science and Electrical Engineering, Beihang University, Xueyuan Street 37, Beijing, China

and more manufacturing enterprises have started to establish their cloud manufacturing platforms [17]. During the running of cloud manufacturing platforms, a large amount of data is generated continuously which is therefore called manufacturing big data. In order to process the big data and improve the capability of big services in manufacturing environment, many mathematical methods [30] and optimization models [31] have been proposed. However, the big data distributes in each stage in manufacturing process, various stages and belong to many departments. In the factory, with the application of all kinds of management software, the most data in many departments cannot be shared. As the result, many information islands are formed. One of the goals of cloud manufacturing is to bridge the gaps between these information islands, so that all manufacturing stages can coordinate, and achieve the overall optimization and scheduling [13]. The Master Data (MD) is the mapping of the reference data in subsystems, It can describe the business entities and link the cross-system business processes [14]. The application of master data can realise key data sharing and automatic, accurate and timely data distribution, analysis and verification in the scope of the entire enterprise. According to the data characteristics in the environment of cloud manufacturing [18], the method of master data can be used to manage the core data. Through the management of master data, the data of all stages in the cloud manufacturing environment can be opened to realise the complete and rapid implementation of the supply chain both in and out of the factory.

The key function of the Master Data Management (MDM) is the ”Management”. The MDM longitudinal structure does not create a new data or new vertical structure of data. On the contrary, it provides a way for enterprises to effectively manage the data stored in the distribution system. The MDM is built based on existing systems and fetches the latest information from these systems and uses advanced technology and processes for automatic, accurate and timely data distribution, analysis and verification in the scope of the entire enterprise. The MDM has the following capabilities [15]:

- Mining the information, knowledge and potential information of subsystems in the range of applications across the enterprise.
- Collecting data in subsystems and sharing the data as a data service.
- Creating a unified master data structure for customers, products, suppliers.
- Supporting multi-user management of data including the permissions of appending, updating, browsing.
- Integrating the management of product information, customer relationship, customer information and so on, which propose data-oriented solutions.
- Establishing the different data management policies for various departments and businesses.

MDM provides a low-cost maintenance solution for the inheritance and reference of the data. In the MDM system, the dimension data determines the source of the data warehouse and makes the data warehouse focused on data volume management and delivery of data management goals. The MDM provides the following functions:

- The Logic of ”match and merge” from one or more source system which identifies and integrates the duplicate records.
- The broad cell level correlation together with historical records which offer a detailed audit trail data content.
- Application of all data sources and applications across all the database in the middle of the relational data.

These functions will significantly reduce the cost of development and maintenance related to the data warehouse. The implement of MDM system requires an architecture suitable for the application environment, a set of tools and a set of strict rules. In this way, the master data can be effectively managed, but there are still a lot of problems in the maintenance process. Master data is established by multiple departments and is constantly maintained and updated in the business application process. This process is difficult to guarantee the data quality of master data and the poor quality of the master data has a significant negative effect on an enterprise [6]. Therefore, an MDM system needs an evaluation method for master data, especially for the data network composed of master data. In this paper, a model of the master data network is built based on the theory of Set Pair Analysis (SPA). In order to verify the master data, an evaluation method for the network is proposed. There are several features with the proposed model and method. Firstly, connection numbers are used to define the uncertainty relationships between data to establish an uncertainty relation matrix. Secondly, the common tendency of the data network is obtained by the analysis of the set-pair tendency for the evaluation of the data network.

The remainder of this paper is organized as follows. The method of master data management and the concept of SPA are reviewed in Section 2. The architecture of master data management system is presented in Section 3. Then, based on the architecture and theory of SPA the methods of modelling and evaluation are presented in Section 4. In Section 5, a case study of an enterprise is provided to represent the methods of modelling with an evaluation which can describe and assess the data network effectively. Finally, Section 6 concludes the paper.

2 Related works

In recent years, there have been a lot of researches on social network services, user correlations [36, 37], user behavior [38], which involve many researches on social networks. However, the researches on data networks formed by manufacturing data are very limited. With the developing of big data, the research on master data is growing. Among of them, the main researches are topics of framework and application.

2.1 Management of master data

A service-oriented architecture for master data exchange is presented, which support requirements described in different parts of the standard, such as developing a data dictionary of master data terms [20]. A knowledge-based scheduling approach of machine considering the uncertainty of master data is proposed [5]. The approach uses the production-condition dependent execution time of all throughput time components with consideration of uncertainty to enable the schedule. For the problem of data cleaning, a data cleaning system (KATA-RA) is presented [2]. It is a knowledge based data cleaning system by crowdsourcing. It can align the semantics with the knowledge database in condition with given a table, a knowledge database, and a crowd, and identify correct and incorrect data, and generate top-k possible repairs for incorrect data. In addition, in the application of knowledge model, a mathematical model is presented, the model can find out the optimal assembly sequence tree, which based on an existing product family within the assembly sequence trees of individual product family members [10]. By the experience with cutting-edge projects, an SOA approach for managing master data of enterprise is presented [3]. The study introduces MDM patterns, blueprints, solutions, best practices. Data mining and knowledge discovery in the context of big data can be enabled through the semantic Web [19]. And there are many different approaches in different stages of the process of knowledge discovery. In addition, The developing of cloud technology also provides more framework support for master data management. In the cloud-based framework, the cloud plane is used to process the data of large-scale, long-term and global, which needs the high-performance processing of master data. Therefore, a tensor-based big service framework is presented [27], which includes a sensing plane, a cloud plane and an application plane. Under this background, a tensor-based cloud-edge computing framework and a tensor-based big data-driven routing recommendation approach are presented [25, 26]. This indicates that the application of master data tends to the cloud. Based on the researches above about MDM, the main researches of master data management focus on the framework, development method, maintenance method and so on.

2.2 Evaluation of data network

There have many researches about architecture of master data management, many topics have been discussed in depth such as data extraction, sharing strategy, linkage mechanism, data description method. Discussion of evaluation methods for master data is insufficient. Master data enables the distributed data to form a network, and the evaluation of master data can refer to the evaluation methods of some other networks. For example, social networks are more focused on trust. A modified flow-based trust evaluation using network flow is presented, to model trust decay with the leakage associated with each node [8]. A graph-based trust evaluation models are presented, and is used in online social networks [9]. For the Internet of Things (IoT), an IoTrust is presented, which is a trust architecture that integrated Soft Defined Network (SDN) in IoT [1]. For the social network, an approach is proposed, which enables the seeking of information and sharing of knowledge in the same networking environments based on the user-generated data and social behaviors [40]. For the modeling and analysis of multi-academic network, a multidimensional network model is presented, which can describe and quantify the multi-type relationships among the academic entities [39]. The data in the data network comes from the distributed system. Besides the trust of data, we should also pay more focus to the network structure, in order to verify whether the master data can represent the trend of distributed data. Studying the evaluation of data network is also to improve the capability of management of master data. There is not much research on the master data network, but many studies on the evaluation of big data. The evaluation of big data main focuses on: execution time, scalability, resource utilization, energy efficiency, microarchitectural behavior, key research fndings and so on [24]. There are many methods can evaluate the performance indexes [4, 22]. But, there are not more evaluation methods for data trends.

However, how to evaluate the data network based on master data is still a problem to be considered in the cloud manufacturing environment. Therefore, an objective evaluation method is needed, to deal with the data network.

2.3 Concept of Set Pair Analysis

In a cloud manufacturing system, various cross-domain manufacturing stage generate different data. There are certain or uncertain relationships among the cross-domain data. For example, the primary key and foreign key between data tables in a system, the cross-system table-to-table relationship linked by master data, the potential cross-system table-to-table relationship indirectly linked by master data in the supply chain. These relationships establish a data network by master data and keys. However, how to evaluate the quality of the data network is an important problem. In this paper, we use the theory of SPA to model the data network and use the connection number to define the relationship between the cross-domain data.

Set Pair Analysis (SPA) is a method of systematic analysis for solving uncertain problems [33, 34] . For the uncertainty of the master data network, the theory of SPA can effectively characterize the master data network.

SPA represents the relationship between the two sets of data. Assuming there is a data set A and a data set B in a complex system, and there is a relationship between A and B. The total number of characteristics between the two groups is defined N. The total number of common characteristics between the two groups is defined S. The total number of opposite characteristics between the two groups is defined P. The total number of uncertain characteristics between the two groups is defined F. The connection number between the two groups is defined u. are as follows:

$$u = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \tag{1}$$

where, $\frac{S}{N} + \frac{F}{N} + \frac{P}{N} = 1$, also expressed as: $\frac{S}{N} = a$, $\frac{F}{N} = b$, $\frac{P}{N} = c$, is as follow:

$$u = a + bi + cj \tag{2}$$

where, $a + b + c = 1$, and i is the uncertainty coefficient, then $i \in [-1, 1]$, the value is defined according to the different application environment. j is the opposite coefficient, for the most case, it is defined the mark value, the value is determined -1 generally. The SPA can describe the uncertainty relationship between two groups of set, and we use the connection number to express the degree of correlation between the two sets.

According to the environment, the Eq. 2 can be transformed to:

$$u' = a + bi \tag{3}$$

$$u'' = a + cj \tag{4}$$

$$u''' = bi + cj \tag{5}$$

where, u' is the relation of similarity and exclusivity, u'' is the relation of similarity and reversity, u''' is the relation of exclusivity and reversity.

In the cloud manufacturing system, the connectivity between data has a certain direction. The change of data set a can cause the change of data set b. The data set b may be linked to data set c.

There are many cases in which SPA is used for evaluation. A dynamic prediction model based on SPA is presented to predict the growth tendency of integrated carrying capacity [28]. A model integrated the theories of SPA and Markov Chain is proposed to evaluate the groundwater quality in a city [23]. A case of ecological evaluation of the competitiveness, using the SPA theory, is studied [29]. But, it lacks in the manufacturing field that the application based on SPA theory.

Therefore, SPA can not only build a model of data network, but its characteristics can also be used to evaluate trends in distributed data.

3 Architecture of the master data management

In the cloud manufacturing system, the application challenges such as long-distance storage of the whole chain information of the supply chain, different systems and large gap in data structure make it difficult to quickly, effectively and safely obtain the collaborative data of the links of the supply chain. The goal of architecture of the master data management is controlling of sharing, access, distribution of data [16]. Based on that, the architecture of MDM should establish the relationship of data and ensure the quality of data.

In this paper, the master data management system realizes the management of cross-domain data for supply chain. The key data of distributed data in various stages and domains in the manufacturing environment are extracted, and these data are formed into master data, and the cloud master database is established. The master data management system establishes many management functions in the cloud. The functions realize data extraction, data modeling, data distribution, data governance, and other functions. Finally, it provides supply-chain-oriented cloud services such as extraction, cleaning, tracing and security control of cross-domain heterogeneous data.

As shown in Figure 1, an architecture of master data management divides four layers, include distributed data resource, cloud master data, middleware, supply chain oriented cross-domain application.

3.1 Distributed Data Resource Layer

Distributed data refers to in the process of processing and manufacturing, the distributed departments, distributed domain, distributed equipment to produce all kinds of data. All information related to manufacturing are categorized and stored, include drawings, documents, cases, guides and so on. The information and data collected from the whole cloud or other clouds are stored in the databases by automatic collection or actively push. The databases are distributed to different areas by different data storage modes.

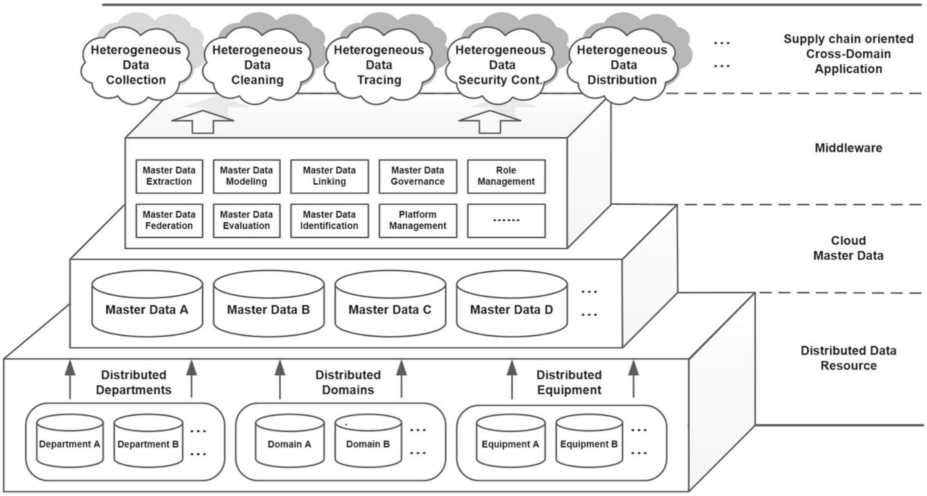

Figure 1 Architecture of Master data Management

3.2 Cloud master data layer

Cloud master data is the mean that establishes many master database in the cloud. The master database represent the share data in different task dimension of database. Master data management make the manufacturing enterprises can be rapidly deployed and easily extend, in order to solve the problem of multiple and distributed by the department of business. MDM can pass 5 steps in the field of intelligent manufacturing environment more master data management:

- **Modelling**: Using the flexible data model to define any type of master data.
- **Identification**: Fast matching and accurately identify repeat project.
- **Federation**: Providing a single virtual view of master data, including from one source to one goal system or from multi-sources to multi-goal systems.
- **Linking**: Revealing the relationship between the various types of master data.
- **Governance**: Creating, using, managment and monitoring the data.

MDM provides business users and data administrator access to powerful interface, so as to realize complete data management and data exception handling, on the other hand, can show different multi-level structure of master data entities.

3.3 Middleware layer

The middleware layer provides the role of managing the bottom-level data and supporting the top-level applications for the master data management system. This includes the extraction, modeling, identification, connection, federation, governance and evaluation of master data, service encapsulation, service search and service combination for top-level cloud service applications, and role management, platform management and computing resource management for the platform itself.

3.4 Supply chain oriented cross-domain application layer

The top layer of the master data management system provides cross-domain cloud applications to the entire supply chain. In the cloud manufacturing environment, the collaborative data of link of the supply chain are all heterogeneous data from different data sources in various domains. Therefore this layer provides much application such as heterogeneous data collection, heterogeneous data cleaning, heterogeneous data tracing, heterogeneous data security control, heterogeneous data distribution, etc.

By the disscussion above, the master data management system can solve the problems of cross-domain acquisition of heterogeneous data in the supply chain, intelligent fusion and efficient control, security sharing and isolation.

4 Modeling and evaluation of data network

In a cloud manufacturing system, we assume that each department represents a manufacturing stage. Each department has an independent management subsystem. The data in the subsystem has certain correlation relation. Through the introduction of Section 3, MDM system establishes the management of master data in the cloud and realizes the mapping of master data in each subsystem. Thus the data in each subsystem form a cross-system correlation. Therefore, the data in the cloud manufacturing system establishes a data network through master data and data keys. As shown in Figure 2.

4.1 Modeling of data network

The relation among data in data network is master data and data key. According to the SPA theory, the data network model is determined as follows: We define two subsystem data are DA_i and DA_j, the set of the data table is as follow:

$$I_{DA_i} = \{I_{i_1}, I_{i_2}...I_{i_k}\}, I_{DA_j} = \{I_{j_1}, I_{j_2}...I_{j_l}\} \tag{6}$$

The total number of tables of subsystem DA_i is $Sum(I_{DA_i}) = k$, the total number of tables DA_j is $Sum(I_{DA_j}) = l$, and the tota numberl of tables is as follows:

$$N = Sum(I_{DA_i} \cup I_{DA_j}),\ \text{(k or l)} \leq N \leq (k + l) \tag{7}$$

Through the extraction of master data, we can determine the explicit correlation between data and data. However, there are many potential relationships, such as the quotation of material in the out purchase stage, which may affect the product selection in the design stage or the supplier selection in the assembling stage. In addition, there are many completely unrelated data relationships, such as the name of the material and the address of the invoice.

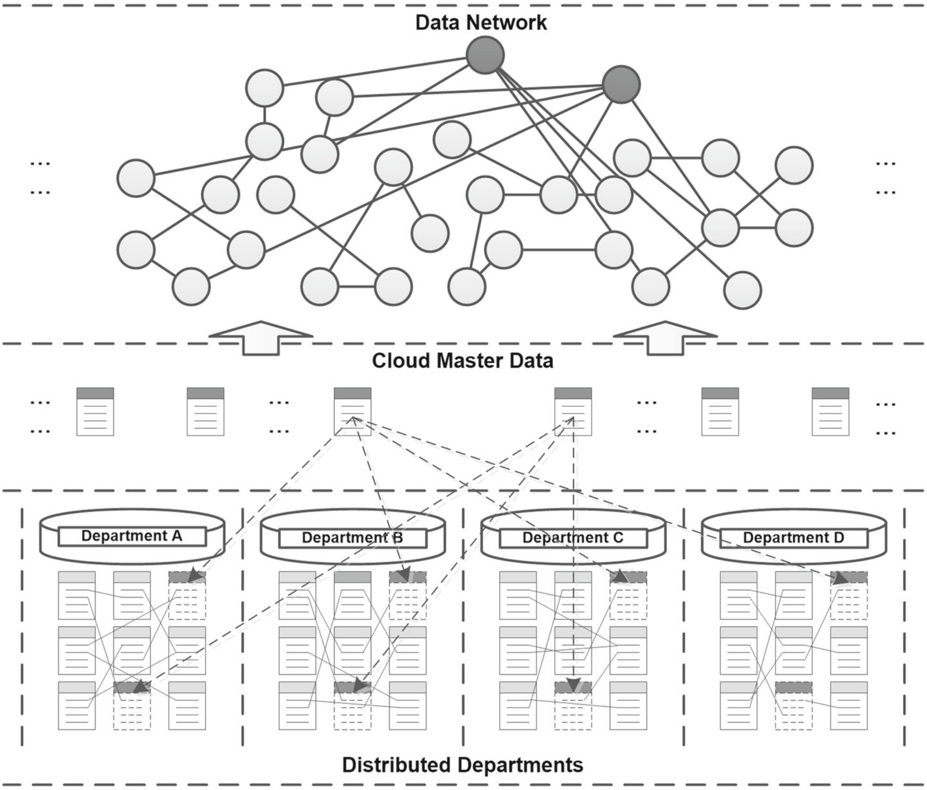

Figure 2 Establishment of Data Network

With these parameters, we can build a model of data network based on the connection number.

For two different stages of data table, the total number of the data column is defined as N, the total number of the common or related data column is defined as S, the total number of the unrelated data column is defined as P, the total number of the potential or uncertained related data column is defined as F. Therefore, for comparing two sets of data, when $I_{DA_i} = I_{DA_j}$, the two sets data is positively correlated, and $S = N$, $F = 0$, $P = 0$. In the most conditions, there is an uncertain relationship between the two sets of data, is defined as:$I_{DA_i} \approx I_{DA_j}$, is as follows:

$$S = Sum(I_{DA_i} \cap I_{DA_j}),\ F = N - Sum(I_{DA_i} \cap I_{DA_j}),\ P = 0 \tag{8}$$

$$S = Sum(I_{DA_i} \cap I_{DA_j}),\ F = 0,\ P = N - Sum(I_{DA_i} \cap I_{DA_j}) \tag{9}$$

$$S = Sum(I_{DA_i} \cap I_{DA_j}),\ F = Sum(I_{DA_i} \approx I_{DA_j}),\ P = N - F - Sum(I_{DA_i} \cap I_{DA_j}) \tag{10}$$

The two sets of data is as follows:

$$a_{(DA_i,DA_j)} = \frac{S}{N},\ b_{(DA_i,DA_j)} = \frac{F}{N},\ c_{(DA_i,DA_j)} = \frac{P}{N} \tag{11}$$

The connection number $u_{(i,j)}$ of two sets of data is as follow:

$$u_{(DA_i,DA_j)} = a_{(DA_i,DA_j)} + b_{(DA_i,DA_j)}i + c_{(DA_i,DA_j)}j \tag{12}$$

which $i \in [-1, 1]$, $j = -1$, $a_{(DA_i,DA_j)} + b_{(DA_i,DA_j)} + c_{(DA_i,DA_j)} = 1$.

Assuming a cloud manufacturing system includes n cross-domain heterogeneous data table, the SPA matrix of the data network is defined $R_{n\times n}$, is as follows:

$$R_{n\times n} = \begin{bmatrix} 0 & u_{(DA_1,DA_2)} & \cdots & u_{(DA_1,DA_n)} \\ u_{(DA_2,DA_1)} & 0 & & u_{(DA_2,DA_n)} \\ \vdots & & \ddots & \vdots \\ u_{(DA_n,DA_1)} & u_{(DA_n,DA_2)} & \cdots & 0 \end{bmatrix} \tag{13}$$

where, the R matrix presents the relationship among the data in data network.

4.2 Evaluation of data network

By the introduction of Section 4.1, a model of the data network is built. However, how to evaluate the rationality of master data is an important problem. The master data is used to build the relationship of data in data network. So the evaluation of the data network is necessary. In this paper, the common tendency is used to evaluate the network as a factor, is as follows:

$$shi(H) = \frac{a_{(DA_i,DA_j)}}{c_{(DA_i,DA_j)}} \tag{14}$$

In the (14), when $c_{(DA_i,DA_j)} \neq 0$, the specific value of $a_{(DA_i,DA_j)}$ and $c_{(DA_i,DA_j)}$ is common tendency of data network. If $c_{(DA_i,DA_j)} \neq 0$, it means there are uncertain relationships between two sets of data. If $\frac{a_{(DA_i,DA_j)}}{c_{(DA_i,DA_j)}} > 1$, it means that the similarity of two sets of data is more than reversity, which can be defined as the common tendency, as follow:

$$shi(H)_s = \frac{a_{(DA_i,DA_j)}}{c_{(DA_i,DA_j)}}, a > c \tag{15}$$

Table 1 Five types of common tendency

Equations	Conditions
$shi(H)_{sS}$	$a > c > b$
$shi(H)_{sW}$	$a > b > c$
$shi(H)_{sM}$	$b > a > c$
$shi(H)_{sQ}$	$a > c, b \to 0$
$shi(H)_{sF}$	$c \to 0, b \to 0$

In a cloud manufacturing system, the more common related between two sets of cross-system data, the closer the relationship between them. In this work, five types of common tendency are presented, as shown in Table 1.

As Shown in Table 1, $shi(H)_{sS}$ is defined as the relationship between the two sets of data with domination by the common tendency, which its degree of interaction is very high in a data network. $shi(H)_{sW}$ is defined as the relationship between the two sets of data with a weak tendency, Although the trend is the same, there are more uncertainties and therefore relatively weak. For the data network, these two sets of data are more uncertainty. $shi(H)_{sM}$ is defined as the relationship between two sets of data with is a very weak tendency, there are many uncertainties and uncertainty of data network is very much. For the data network, potential relationships are the main characteristics of the network. $shi(H)_{sQ}$ is defined as the relationship between two sets of data with the uncertainty closes to zero. The relationship between two sets of data is clear. there is not potential relation. $shi(H)_{sF}$ is defined as the relationship between the two sets of data with completely relation, both uncertain and unrelated data are close to zero. The two sets of data describe the same things mostly. In the cloud manufacturing environment, this condition may be that it is a mapping of master data in two stages. According to (15), the common tendency matrix is as follow:

$$R_{n\times n}(shi(H)_s) = \begin{bmatrix} 0 & \frac{a_{(DA_1,DA_2)}}{c_{(DA_1,DA_2)}} & \cdots & \frac{a_{(DA_1,DA_n)}}{c_{(DA_1,DA_n)}} \\ \frac{a_{(DA_2,DA_1)}}{c_{(DA_2,DA_1)}} & 0 & & \frac{a_{(DA_2,DA_n)}}{c_{(DA_2,DA_n)}} \\ \vdots & & \ddots & \vdots \\ \frac{a_{(DA_n,DA_1)}}{c_{(DA_n,DA_1)}} & \frac{a_{(DA_n,DA_2)}}{c_{(DA_n,DA_2)}} & \cdots & 0 \end{bmatrix} \tag{16}$$

According to the matrix, the average of common tendency of $R_{n\times n}(shi(H)_s)$ is defined as $C(shi(H)_s)$, is as follows:

$$C(shi(H)_s) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_{(DA_i,DA_j)}}{c_{(DA_i,DA_j)}}}{n\times n} \tag{17}$$

The average common tendency shows the overall tendency of the entire data network and can enable verification of the overall connection tendency of the data network. $shi(H)_{sS}$ is defined as a strong average common tendency of the data network, which most of data in the data network can build the relationship and most of the subsystem can involve in the supply chain. Both of $shi(H)_{sW}$ and $shi(H)_{sM}$ are defined as the weaker average common tendency of the data network, which the connection number of data in the network is not high and many data in the subsystem only relate to own subsystem. The subsystems are isolated. The uncertain relation of $shi(H)_{sM}$ is more than $shi(H)_{sW}$. $shi(H)_{sQ}$ and $shi(H)_{sF}$ are defined as the certain average common tendency, which almost no uncertain relationship in the data network. This is an ideal situation where all the subsystems in the system follow the same data specification and rule, and the subsystems are completely transparent and shared. $shi(H)_{sF}$ means that the common tendency is completely the same. It could be that the two subsystems are exactly the same.

5 Case study

In this section, a case study from an aviation product enterprise is presented to represent the modeling and evaluation of the data network by SPA.

In the manufacturing environment, three typical manufacturing processes are selected, including out purchase, outsourcing production, inter fine production. This system includes three subsystems, which 35 data tables with strong relationship with foreign keys and master data are selected from the data of three systems, and some data tables uncross-system are removed, such as invoices address, classification of vender, pictures and dictionaries.

As shown in Figure 3, a cross-system data network is established. The connection number matrix among data is as follows:

$$R_{35\times 35}=\begin{bmatrix} 0 & 0.5+0.2i+0.3j & 0.7+0.1i+0.2j & \cdots & 0.7+0.0i+0.3j \\ 0.5+0.2i+0.3j & 0 & 0.4+0.1i+0.5j & \cdots & 0.3+0.0i+0.7j \\ 0.7+0.1i+0.2j & 0.4+0.1i+0.5j & 0 & \cdots & 0.6+0.2i+0.2j \\ & \vdots & \vdots & \ddots & \vdots \\ 0.7+0.1i+0.2j & 0.4+0.2i+0.4j & 0.5+0.2i+0.3j & \cdots & \\ 0.7+0.0i+0.3j & 0.3+0.0i+0.7j & 0.6+0.2i+0.2j & \cdots & 0 \end{bmatrix} \tag{18}$$

As shown in (18), the connection number matrix $R_{35\times 35}$ among data is calculated through the relationship of master data in the data network. The properties linked by master data or data key are $a_{(DA_i,DA_j)}$. The properties linked by potential relation are $b_{(DA_i,DA_j)}$.

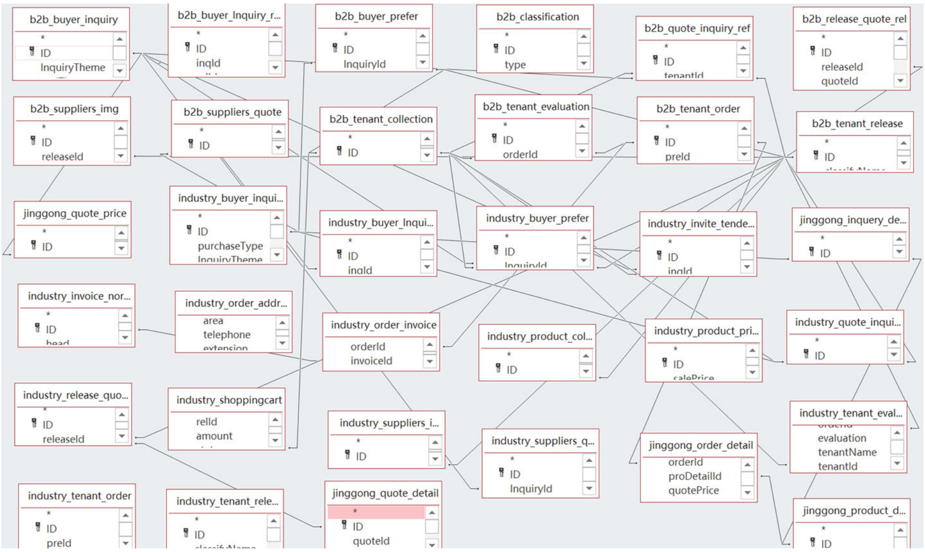

Figure 3 Cross-system Data Network

The non-relationship properties are $c_{(DA_i,DA_j)}$. Thus, the common tendency matrix is calculated as:

$$R_{35\times 35}(shi(H)_s) = \begin{bmatrix} 0 & 1.55 & 3.5 & 1.25 & \cdots & 3.5 & 2.33 \\ 1.55 & 0 & 0.8 & 7.0 & \cdots & 1 & 0.42 \\ 3.5 & 0.8 & 0 & 1.25 & \cdots & 1.66 & 3 \\ & & \vdots & & \ddots & \vdots & \vdots \\ 3.5 & 1 & 1.66 & & \cdots & & \\ 2.33 & 0.42 & 3 & & \cdots & & 0 \end{bmatrix} \tag{19}$$

Table 2 Average common tendencies of all elements

$C(u_{(DA_1,DA_k)})$	$shi(H)_{sS} = 1.76$	$0.49 + 0.10i + 0.41j$
$C(u_{(DA_2,DA_k)})$	$shi(H)_{sS} = 1.94$	$0.50 + 0.10i + 0.40j$
$C(u_{(DA_3,DA_k)})$	$shi(H)_{sS} = 1.43$	$0.49 + 0.06i + 0.44j$
$C(u_{(DA_4,DA_k)})$	$shi(H)_{sS} = 2.04$	$0.51 + 0.09i + 0.40j$
$C(u_{(DA_5,DA_k)})$	$shi(H)_{sS} = 1.97$	$0.49 + 0.13i + 0.38j$
$C(u_{(DA_6,DA_k)})$	$shi(H)_{sS} = 2.30$	$0.53 + 0.12i + 0.35j$
$C(u_{(DA_7,DA_k)})$	$shi(H)_{sS} = 1.92$	$0.52 + 0.10i + 0.38j$
$C(u_{(DA_8,DA_k)})$	$shi(H)_{sS} = 1.90$	$0.50 + 0.10i + 0.40j$
$C(u_{(DA_9,DA_k)})$	$shi(H)_{sS} = 1.84$	$0.53 + 0.10i + 0.37j$
$C(u_{(DA_{10},DA_k)})$	$shi(H)_{sS} = 1.25$	$0.45 + 0.11i + 0.44j$
$C(u_{(DA_{11},DA_k)})$	$shi(H)_{sS} = 2.15$	$0.52 + 0.11i + 0.37j$
$C(u_{(DA_{12},DA_k)})$	$shi(H)_{sS} = 1.55$	$0.51 + 0.07i + 0.42j$
$C(u_{(DA_{13},DA_k)})$	$shi(H)_{sS} = 1.84$	$0.51 + 0.09i + 0.39j$
$C(u_{(DA_{14},DA_k)})$	$shi(H)_{sS} = 1.79$	$0.49 + 0.09i + 0.43j$
$C(u_{(DA_{15},DA_k)})$	$shi(H)_{sS} = 1.83$	$0.51 + 0.08i + 0.42j$
$C(u_{(DA_{16},DA_k)})$	$shi(H)_{sS} = 2.01$	$0.51 + 0.11i + 0.38j$
$C(u_{(DA_{17},DA_k)})$	$shi(H)_{sS} = 1.50$	$0.49 + 0.12i + 0.39j$
$C(u_{(DA_{18},DA_k)})$	$shi(H)_{sS} = 1.88$	$0.50 + 0.09i + 0.41j$
$C(u_{(DA_{19},DA_k)})$	$shi(H)_{sS} = 1.50$	$0.49 + 0.12i + 0.39j$
$C(u_{(DA_{20},DA_k)})$	$shi(H)_{sS} = 1.73$	$0.50 + 0.11i + 0.39j$
$C(u_{(DA_{21},DA_k)})$	$shi(H)_{sS} = 2.11$	$0.50 + 0.12i + 0.38j$
$C(u_{(DA_{22},DA_k)})$	$shi(H)_{sS} = 2.01$	$0.49 + 0.09i + 0.41j$
$C(u_{(DA_{23},DA_k)})$	$shi(H)_{sS} = 1.91$	$0.51 + 0.10i + 0.39j$
$C(u_{(DA_{24},DA_k)})$	$shi(H)_{sS} = 1.97$	$0.50 + 0.11i + 0.39j$
$C(u_{(DA_{25},DA_k)})$	$shi(H)_{sS} = 2.07$	$0.52 + 0.10i + 0.38j$
$C(u_{(DA_{26},DA_k)})$	$shi(H)_{sS} = 1.67$	$0.49 + 0.13i + 0.38j$
$C(u_{(DA_{27},DA_k)})$	$shi(H)_{sS} = 2.05$	$0.51 + 0.07i + 0.41j$
$C(u_{(DA_{28},DA_k)})$	$shi(H)_{sS} = 2.18$	$0.54 + 0.11i + 0.35j$
$C(u_{(DA_{29},DA_k)})$	$shi(H)_{sS} = 1.90$	$0.51 + 0.13i + 0.35j$
$C(u_{(DA_{30},DA_k)})$	$shi(H)_{sS} = 1.97$	$0.50 + 0.12i + 0.38j$
$C(u_{(DA_{31},DA_k)})$	$shi(H)_{sS} = 1.87$	$0.54 + 0.09i + 0.37j$
$C(u_{(DA_{32},DA_k)})$	$shi(H)_{sS} = 2.19$	$0.50 + 0.12i + 0.38j$
$C(u_{(DA_{33},DA_k)})$	$shi(H)_{sS} = 1.95$	$0.51 + 0.09i + 0.40j$
$C(u_{(DA_{34},DA_k)})$	$shi(H)_{sS} = 1.92$	$0.49 + 0.12i + 0.39j$
$C(u_{(DA_{35},DA_k)})$	$shi(H)_{sS} = 1.35$	$0.50 + 0.08i + 0.42j$

As shown in (18) and (19), the diagonal elements in the connection number matrix are 0, which are the connection number between a data and itself, so no calculation is done here. Where, the connection number is less than zero, such as: $a_{(DA_i,DA_j)} < c_{(DA_i,DA_j)}$, indicating that the connection number of two sets of data is low and there are no relation between them.

$$C(u_{(DA_c,DA_k)}) = \frac{\sum_{k=1}^{35} a_{(DA_c,DA_k)}}{n-1} + \frac{\sum_{k=1}^{35} b_{(DA_c,DA_k)}}{n-1} i + \frac{\sum_{k=1}^{35} c_{(DA_c,DA_k)}}{n-1} j \tag{20}$$

where, $c, k \in [1, n], c \neq k, n = 35$.

As shown in (20), $C(u_{(DA_c,DA_k)})$ is the average common tendency between each data and other data. And the average common tendencies of all elements are as shown in Table 2.

As shown in Table 2, in the data network, the average common tendencies of the data other data are all strong, indicating that the master data of the system is established reasonably. The data network of this system can make the most of key data have a strong common tendency with other data. Similarly, the average common tendency of the entire data network is as follow:

$$C(shi(H)_{sS}(DA_i, DA_j)) = 1.864 \tag{21}$$

It can be shown in (21) that the average common tendency of the data network is more than 1, and the common tendency is strong, and the uncertainty is less than the certainty. The overall connection number of data network tends to be the same trend, with less uncertainty, the relationship among the cross-system data is clear.

6 Conclusion

This paper first introduces the concept of SPA. Then, an architecture of master data management is presented. It is the basis of the master data system in this paper. Through the MDM system, a data network based on master data and data key can be established. Using the SPA with common tendency parameters, this work forms an evaluation standard of the network. The common tendency can evaluate the tendency of master data network by evaluating the average common tendency between data in a data network. Common tendency affects master data networks differently change in different business models. In the cloud manufacturing environment, the timely update, response, distribution of data in each stage are important factors to ensure the efficient operation of the supply chain. Through the evaluation of the data network, the master data can be adjusted in real time and its potential correlation can be found.

In the future, we will focus on the knowledge-based master data management and build a high-dimensional mapping of master data through knowledge model. Finding the potential relationship between data and data in the high-dimensional space and updating the master data in real time.

Acknowledgments This research is supported by the National Key Research and Development Program of China No. 2018YFB1004001, and the NSFC (National Science Foundation of China) project No. 61572057 and 61836001, and the Scientific Research Fund Project of Beijing Information Science & Technology University No.5221810921.

References

1. Chen, J., Tian, Z., Cui, X., Yin, L., Wang, X.: Trust architecture and reputation evaluation for internet of things. Journal of Ambient Intelligence and Humanized Computing: 1–9. https://doi.org/10.1007/s12652-018-0887-z (2018)
2. Chu, X., Morcos, J., IF Ilyas, M., Ouzzani, P., Katara, P., et al.: A data cleaning system powered by knowledge bases and crowdsourcing. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data ACM. https://doi.org/10.1145/2723372.2749431 (2015)
3. Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., et al.: Enterprise Master Data Management (Paperback): An SOA Approach to Managing Core Information. Pearson Education (2008)
4. Gani, A., Siddiqa, A., Shamshirband, S., et al.: A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowl. Inf. Syst. **46.2**, 241–284 (2016). https://doi.org/10.1007/s10115-0830-y
5. Geiger, F., Reinhart, G.: Knowledge-based machine scheduling under consideration of uncertainties in master data. Prod. Eng. **10.2**, 197–207 (2016). https://doi.org/10.1007/s11740-015-0652-5
6. Haug, A., Arlbjørn, J.S.: Barriers to master data quality. J. Enter. Inf. Manag. **24**(3), 288–303 (2011). https://doi.org/10.1108/17410391111122862
7. Jeschke, S., Brecher, C., Meisen, T., et al.: Industrial internet of things and cyber manufacturing systems. Industrial Internet of Things, pp. 3–19. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-42559-7_1
8. Jiang, W., Wu, J., Li, F., Wang, G., et al.: Trust evaluation in online social networks using generalized network flow. IEEE Trans. Comput. **65.3**, 952–963 (2016). https://doi.org/10.1109/TC.2015.2435785
9. Jiang, W., Wang, G., Bhuiyan, M., Wu, J.: Understanding graph-based trust evaluation in online social networks: Methodologies and challenges. ACM Comput. Surv. (CSUR) **49.1**, 10 (2016). https://doi.org/10.1145/2906151
10. Kashkoush, M., ElMaraghy, H.: Knowledge-based model for constructing master assembly sequence. J. Manuf. Syst. **34**, 43–52 (2015). https://doi.org/10.1016/j.jmsy.2014.10.004
11. Lasi, H., Fettke, P., Kemper, H., et al.: Industry 4.0. Bus. Inf. Syst. Eng. **6.4**, 239–242 (2014). https://doi.org/10.1109/THMS.2017.2725341
12. Li, B.H., Zhang, L., Wang, S.L., Tao, F., Cao, J., et al.: Cloud manufacturing: a new service-oriented networked manufacturing model. Comput. Integr. Manuf. Syst. **16.1**, 1–7 (2010)
13. Li, B.H., Zhang, L., Ren, L., Chai, X.D., Tao, F., et al.: Typical characteristics, technologies and applications of cloud manufacturing. Comput. Integr. Manuf. Syst. **18.7**, 1345–1356 (2012)
14. Linstedt, D., Olschimke, M.: Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann, San Mateo (2015)
15. Loshin, D.: Master data management. Morgan Kaufmann, San Mateo (2010)
16. Otto, B.: How to design the master data architecture: Findings from a case study at Bosch. Int. J. Inf. Manag. **32.4**, 337–346 (2012). https://doi.org/10.1016/j.ijinfomgt.2011.11.018
17. Ren, L., Zhang, L., Tao, F., Zhao, C., Chai, X.D., Zhao, X.P.: Cloud manufacturing: from concept to practice. Enter. Inf. Syst. **9:2**, 186–209 (2015). https://doi.org/10.1080/17517575.2013.839055
18. Ren, L., Zhang, L., Wang, L.H., Tao, F., Chai, X.D.: Cloud manufacturing: key characteristics and applications. Int. J. Comput. Integr. Manuf. **30:6**, 501–515 (2017). https://doi.org/10.1080/0951192X.2014.902105
19. Ristoski, P., Paulheim, H.: Semantic Web in data mining and knowledge discovery: A comprehensive survey. Web semantics: science, services and agents on the World Wide Web 36. pp. 1–22. https://doi.org/10.1016/j.websem.2016.01.001 (2016)
20. Rivas, B., Merino, J., Caballero, I., Serrano, M., et al.: Towards a service architecture for master data exchange based on ISO 8000 with support to process large datasets. Comput. Stand. Interfaces **54**, 94–104 (2017). https://doi.org/10.1016/j.csi.2016.10.004
21. Sheth, A., Anantharam, P., Henson, C.: Physical-cyber-social computing: an early 21st century approach. IEEE Intell. Syst. **1**, 78–82 (2013). https://doi.org/10.1109/MIS.2013.20
22. Song, M.L., Fisher, R., Wang, J.L., Cui, L.B.: Environmental performance evaluation with big data: Theories and methods. Ann. Oper. Res. **270.1-2**, 459–472 (2018). https://doi.org/10.1007/s10479-016-2158-8
23. Su, F., Wu, J., He, S.: Set pair analysis-Markov chain model for groundwater quality assessment and prediction: A case study of Xi'an city, China. Human and Ecological Risk Assessment: An International Journal. https://doi.org/10.1080/10807039.2019.1568860 (2019)

24. Veiga, J., Expósito, R.R., Pardo, X.C., et al.: Performance evaluation of big data frameworks for large-scale data analytics. 2016 IEEE International Conference on Big Data (Big Data) IEEE. https://doi.org/10.1109/BigData.2016.7840633 (2016)
25. Wang, X., Yang, L., Xie, X., Jin, J., et al.: A cloud-edge computing framework for cyber-physical-social services. IEEE Commun. Mag. **55.11**, 80–85 (2017). https://doi.org/10.1109/MCOM.2017.1700360
26. Wang, X., Yang, L., Kuang, L., Liu, X., Zhang, Q., et al.: A tensor-based big-data-driven routing recommendation approach for heterogeneous networks. IEEE Netw. **33.1**, 64–69 (2018). https://doi.org/10.1109/MNET.2018.1800192
27. Wang, X., Yang, L., Feng, J., Chen, X., et al.: A tensor-based big service framework for enhanced living environments. IEEE Cloud Comput. **3.6**, 36–43 (2016). https://doi.org/10.1109/MCC.2016.130
28. Wei, C., Dai, X., Ye, S., Guo, Z., Wu, J.: Prediction analysis model of integrated carrying capacity using set pair analysis. Ocean Coast. Manag. **120**, 39–48 (2016). https://doi.org/10.1016/j.ocecoaman.2015.11.011
29. Xiao, N.: Research on the ecological evaluation of the competitiveness of based on set pair Analysis-A case study. Chem. Eng. Trans. **51**, 811–816 (2016). https://doi.org/10.3303/CET1651136
30. Yang, L.T., Wang, X., Chen, X., Wang, L., et al.: A multi-order distributed HOSVD with its incremental computing for big services in cyber-physical-social systems. IEEE Transactions on Big Data. https://doi.org/10.1109/TBDATA.2018.2824303 (2018)
31. Yang, L.T., Wang, X., Chen, X., Han, J.J., et al.: A tensor computation and optimization model for cyber-physical-social big data. IEEE Transactions on Sustainable Computing. https://doi.org/10.1109/TSUSC.2017.2777503 (2017)
32. Zhang, L., Luo, Y.L., Tao, F., Li, B.H., Ren, L., Zhang, X.S., Guo, H., Cheng, Y., Hu, A.R., Liu, Y.K.: Cloud manufacturing: a new manufacturing paradigm. Enter. Inf. Syst. **8:2**, 167–187 (2014). https://doi.org/10.1080/17517575.2012.683812
33. Zhao, K.Q.: Set pair analysis and its preliminary application, pp. 1–200. Zhejiang Science and Technology Press, Hangzhou (2000)
34. Zhao, K.Q., Xuan, A.L.: Set pair Theory-A new theory method of Non-Define and its applications [J]. Syst. Eng. **1**, 003 (1996)
35. Zhou, X., Zomaya, A.Y., Li, W., Ruchkin, I.: Cybermatics: Advanced Strategy and Technology for Cyber-Enabled Systems and Applications: 350–353. https://doi.org/10.1016/j.future.2017.09.052 (2018)
36. Zhou, X., Chen, J., Wu, B., Jin, Q.: Discovery of action patterns and user correlations in task-oriented processes for goal-driven learning recommendation. IEEE Trans. Learn. Technol. **7.3**, 231–245 (2014). https://doi.org/10.1109/TLT.2013.2297701
37. Zhou, X., Jin, Q.: A heuristic approach to discovering user correlations from organized social stream data. Multimed. Tools Appl. **76.9**, 11487–11507 (2017). https://doi.org/10.1007/s11042-014-2153-5
38. Zhou, X., Wu, B., Jin, Q.: User role identification based on social behavior and networking analysis for information dissemination. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2017.04.043 (2017)
39. Zhou, X., Liang, W., Wang, K., Huang, R., Jin, Q.: Academic Influence Aware and Multidimensional Network Analysis for Research Collaboration Navigation Based on Scholarly Big Data. In: IEEE Transactions on Emerging Topics in Computing. https://doi.org/10.1109/TETC.2018.2860051
40. Zhou, X., Wu, B., Jin, Q.: Analysis of User Network and Correlation for Community Discovery Based on Topic-Aware Similarity and Behavioral Influence, vol. 48, pp. 559–571 (2018). https://doi.org/10.1109/THMS.2017.2725341

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.