Data Timeliness Evaluation Based on Attribute Value Dependencies

Jun Tang[1,2,4], Bing Guo[1,2], Yan Shen[3], Cheng Dai[1,2], Yuming Jiang[1,2]*

[1] College of Computer Science, Sichuan University, China
[2] Big Data Analysis and Fusion Application Technology Engineering Laboratory of Sichuan Province, China
[3] School of Computer Science, Chengdu University of Information Technology, China
[4] Changhong Central Research Institute, Sichuan Changhong Electronic (Group) Co., Ltd., China
9535390@qq.com, guobing@scu.edu.cn, sheny@cuit.edu.cn, daicheng@scu.edu.cn, jiangym@scu.edu.cn

Abstract

Data timeliness, a critical determinant of data quality, can distort data analysis outcomes and lead decision-makers astray when utilizing outdated data for mining. Current methodologies for exploring data timeliness encompass the recovery of temporal data sequences and the determination of timeliness relying on factors such as timestamps and uncertain data relationships. Timestamp-centric techniques are constrained by their dependence on precise temporal values. In contrast, methods grounded in uncertain relationships among data entities excel in managing multiple tuples of same entity. However, limited research addresses data timeliness while considering dependencies among data attribute values. This paper addresses this gap by investigating data timeliness within the context of attribute value dependencies. The paper commences by introducing pertinent concepts and regulations related to attribute value dependencies. It then defines the prevailing issue and proposes regulations alongside a data timeliness assessment model. Furthermore, the paper presents algorithms for extracting and consolidating data attribute value dependency regulations, in conjunction with an algorithm for timeliness assessment. The efficacy of these algorithms and the accuracy of the data timeliness assessment model are validated using real-world datasets. The experimental results affirm that the proposed algorithms demonstrate high execution efficiency and yield precise data timeliness assessments.

Keywords: Data timeliness, Data quality, Attribute value dependencies, Timeliness assessment

1 Introduction

The transition from the Information Age to the Data Age has marked an era where extracting information from data and mining valuable knowledge to support decision-making becomes pivotal for organizational success [1]. Nonetheless, the big data of many organizations exhibit characteristics such as multi-sourcing, heterogeneity, high dimensionality, high throughput, and rapid data generation. These traits present numerous challenges in data exploration, encompassing data intricacies and latent worth, data segregation and untapped potential, constrained value augmentation, and value hindrance. Consequently, augmenting data quality emerges as a pressing concern in academia.

Scholars in this domain have engaged in data quality control research utilizing methods including direct observation, social surveys, and theoretical deduction. They have recognized accuracy, completeness, consistency, timeliness, and entity identity as paramount evaluation metrics in the data quality control process [2]. Notably in the context of industrial production, the timeliness of data plays a pivotal role in determining industrial production efficiency, safety, and product quality. Failing to promptly and adeptly identify anomalies, faults, and crisis situations linked to equipment, environment, and processes in industrial production can foster latent safety hazards within the production environment. This, in turn, can yield comprehensive losses across the manufacturing system, including product defects, equipment breakdowns, and diminished production capacity. In recent years, the flourishing expansion of the Industrial Internet of Things (IIoT) has spurred industrial digital transformation, introducing a new central production element: data. This data exhibits significant temporal attributes, encompassing timestamps, unique data sources, structured data, and sporadic updates. Moreover, the magnitude of temporal data generated by digital factories is substantial. Investigating the timeliness of industrial data can facilitate intelligent production scheduling, equipment maintenance, and production optimization. Hence, a comprehensive evaluation of data timeliness becomes imperative prior to employing data for guiding industrial production and user service provision.

Within the domain of industrial production, intricate interconnections characterize elements such as materials, equipment, processes, and environment. For example, the execution of orders involves coordinating materials, equipment, and production, which are further supported by inventory and procurement mechanisms. The synchronization of material scheduling, equipment performance, and environmental suitability collectively shapes production advancement and quality. Navigating

*Corresponding Author: Bing Guo; E-mail: guobing@scu.edu.cn
DOI: https://doi.org/10.70003/160792642025032602003

the web of contributing factors influencing changes in production efficiency facilitates the identification of optimal equipment operation parameters and production scheduling. Given the multifaceted interactions among attributes, alterations in these relationships over time reflect the inherent temporal aspect of the data. In real-time analysis of industrial production, the swift generation and application of data result in a limited window of timeliness. For instance, in scenarios entailing real-time monitoring of product quality, timely access to pertinent information impacting product quality—such as production processes, operational parameters, material quality, production efficiency, and continuous production duration—empowers prompt quality assessments. In these production contexts, the significance of data timeliness research is relatively modest, often attainable through predefined time intervals of data. However, in industrial production settings encompassing activities like production planning, inventory management, equipment maintenance, intelligent parameter configuration, and process control, data timeliness critically influences outcome accuracy. Employing outdated data can lead to compromised production efficiency, compromised product quality oversight, skewed factory scheduling, and disorderly material reserves.

Upon this analysis, a pronounced interdependency emerges among diverse data elements within industrial production, fostered by the intricate relationships between entity attributes. However, a review of existing literature on data timeliness assessment underscores the scarcity of studies approaching this subject through the lens of attribute value dependencies. Consequently, this paper undertakes a comprehensive investigation of data timeliness assessment grounded in attribute value dependencies.

2 Literature Review

The matter of data timeliness holds historical significance and currently confronts an exigent state. Scholars have extensively investigated this matter, broadly categorized into two perspectives based on the determinacy of data. Primarily, researchers have delved into data timeliness reliant on precise timestamps [3]. Nonetheless, in practical scenarios, only a fraction of data possesses accurate timestamp values, and rectifying these precise timestamps presents a formidable challenge, thus constricting the applicability of this approach. Secondly, attention has centered on uncertain data attribute relationships. Li et al. devised a derivation algorithm for temporal attribute value relationships grounded in inherent connections among attributes within the same database [4]. This methodology facilitates the evaluation of data obsolescence at a specific juncture, quantitatively elucidating the timeliness of data items, tuples, and datasets. Additionally, Liang et al. amalgamated this temporal recovery algorithm with data mining techniques, specifically decision trees, culminating in a methodology for recuperating student course selection data over time, consequently validating the efficacy of the temporal recovery algorithm [5].

Beyond the exploration of data relationships from a database schema standpoint, scholars have also ventured into data timeliness from the vantage points of data semantics and domain knowledge. Fan et al. tackled missing data value quandaries based on data semantics within an entity comprising multiple tuples [6]. Their inquiry also delved into data timeliness replication, inference, and correlated notions, substantiating the delineation and characterization of temporal correlation. From a domain knowledge angle, Li et al. fused rules and statistics to represent domain knowledge through traditional rules and depicted data's temporal distributional attributes using statistical distributions [7]. This approach aptly fulfilled the prerequisites for timely data assessment in data mining applications. Fan et al. grappled with data timeliness and consistency restoration where the same entity manifests inconsistent attributes among various tuples [8]. They harnessed the local temporal order of data attribute values to convey temporal information, derived temporal constraints to signify data timeliness relationships, and employed constant conditional functional dependencies to ascertain the latest data values. Furthermore, some literature introduced methods augmenting data timeliness through the amalgamation of quality rules and statistical methodologies [9-10]. Some literature introduced techniques encompassing dynamic data sets, including the establishment of entity query B-trees and static/dynamic link lists for entities [10-12]. Others have simultaneously accelerated and compressed convolutional neural networks through multi-stage filtering pruning algorithms to alleviate the dependence on IoT edge node hardware [13] and reduce the high computational cost and huge storage consumption in the cloud. These methodologies advanced the efficiency of data timeliness determination and hastened the updates of data timeliness in voluminous datasets.

These approaches have not only transcended the confines of timestamp-based determinations but also restored temporal sequence relationships in data through domain knowledge, state value transformations, and specific forms of uncertain patterns. Precisely ascertaining the interval range where timestamp information is absent emerges as a rational and potent strategy [10].

As a foundational factor influencing data quality, this paper squarely addresses data timeliness, specifically centering on probing data timeliness assessment rooted in attribute value dependencies. In the context of interrelated attribute relationships, a shift in one attribute's value denotes a corresponding shift in another attribute. This signals that the initial attribute value no longer aligns with contemporary application requisites. Timeliness evaluation of these values markedly mitigates the adverse ramifications stemming from outdated data. The ensuing paper delineates the subsequent research trajectory. First, introduce pertinent rules and definitions, define the quandary, and propose a data timeliness assessment framework. Second, algorithms are presented for the extraction and amalgamation of conditional functional value dependencies, alongside the data timeliness assessment process. Third, substantiate the efficiency and efficacy of the rules and algorithms via real-world production datasets, dissect and deliberate on the authenticity of the data timeliness assessment framework. The paper culminates by succinctly summarizing the research and instigating preliminary dialogues concerning future research trajectories that warrant attention.

3 Data Timeliness Assessment Model

3.1 Concepts and Definitions of Relevant Rules

The definitions presented in this section encompass the essential relationships and characteristics governing attribute values within a relational database schema, particularly focusing on the temporal aspects inherent in these connections over time. Let us denote the relational database schema as R = [*Entity Identifier* (*EID*), A_1, A_2, …, A_n], and tuples with the same *EID* correspond to data from the same entity. We adopt the approaches from references for entity identification [14-15]. A_i represents the i-th attribute of a specific entity, with its scope being *dom* (A_i).

Meanwhile, *dom* (A_i) consists of different temporal values, where each attribute value represents different production conditions, different attribute dependencies, and different degrees of dependency influence. During the data generation process, the attribute values change throughout the attribute's lifecycle.

Definition 1 (Attribute Value Dependency, AVD): Attribute Value Dependency pertains to the consistent underpinning that dictates the interconnections between associated data attribute values within a dataset. Among data tuples of identical entity within R, certain attributes exhibit correlations, occasionally robust correlations. In such instances, values of certain attributes are influenced or directly determined by values of related attributes. We denote these attributes as engaged in an attribute value dependency relationship, formalized as:

$$\phi\text{: } \forall t \in R, X \to Y \tag{1}$$

Here, for a tuple within relation schema R, the attribute values in attribute set Y depend on the attribute values in attribute set X. In this context, X is termed the dependent attribute set of tuple t, while Y is known as the depending attribute set of tuple t, with $X \cap Y = \emptyset$.

Definition 2 (Conditional Attribute Value Dependency, CAVD): Conditional Attribute Value Dependency, within the framework of a relational database schema, characterizes the connection where, subject to specific conditions satisfied by attributes of certain tuple components, the values of certain attributes hinge on the values of other attributes. This relationship is denoted as:

$$\phi : \forall t \in R, (C \subseteq C_0 \middle| X \to Y) \tag{2}$$

For a tuple t in relation schema R, if the values in conditional attribute set C fulfill the domain C_0, then the attribute values within attribute set Y depend on the attribute values within attribute set X. Here, C signifies the conditional attribute set, while C_0 represents the domain set of the conditional attribute set C, with $C \cap X \cap Y = \emptyset$.

Definition 3 (Currency Constraints, CC): In the relational database schema R, for the same entity, if tuple t_2 is temporally more current than tuple t_1, the currency constraint dictates that each attribute value within tuple t_2 is more current than the corresponding attribute value within tuple t_1. Formally:

$$\begin{aligned}&\phi : \forall t_1, t_2 \in R, (t_1[EID] = t_2[EID] \wedge t_1 < t_2)\\ &\to A_i^{t_1} < A_i^{t_2}\end{aligned} \tag{3}$$

Conversely, for any two tuples of the same entity, if the attribute value of A_i within tuple t_2 is temporally more recent than the corresponding attribute value within tuple t_1, the currency constraint mandates that the currency of tuple t_2 surpasses that of tuple t_1. This relationship is expressed as:

$$\begin{aligned}&\phi : \forall t_1, t_2 \in R, (t_1[EID] = t_2[EID] \wedge A_i^{t_1} < A_i^{t_2})\\ &\to t_1 < t_2\end{aligned} \tag{4}$$

3.2 Problem Definition

Following the exposition of attribute value dependency and currency constraints definitions, this paper addresses a problem defined as follows: Within a relational database schema R featuring associated attributes and an instance I , accompanied by attribute value dependency directions from domain experts or experiential knowledge, denoted as dependent attribute set X and depending attribute set Y, the task is to determine the data timeliness of instance I based on attribute value dependency regulations within R and the data timeliness rules outlined in Definition 4.

Definition 4 (Data Timeliness Assessment Rule 1, DTAR1): In the relational database schema R, when considering tuples of the same entity containing associated or strongly correlated attributes within attribute sets X and Y, if the currency of attribute values X in t_2 surpasses that of t_1, then the currency of attribute values Y in t_2 likewise exceeds that of t_1. This rule is represented as:

$$\begin{aligned}&\forall t_1, t_2 \in R, (\exists X, Y \in R \wedge t_1[EID] = t_2[EID]\\ &\wedge X^{t_1} < X^{t_2}) \to Y^{t_1} < Y^{t_2}\end{aligned} \tag{5}$$

Proof: As the entity adheres to the attribute value dependency rule ϕ, the values within attribute set Y for all entity tuples depend on the values within attribute set X. Consequently, changes in Y coincide with changes in X . Given the temporal precedence of attribute value X_i in tuple t_2 over the corresponding value in tuple t_1, based on timeliness rule (4), tuple t_2 surpasses tuple t_1 in timeliness. This deduction stems from timeliness rule (3), affirming that the values within attribute set Y in tuple t_2 surpass the temporal currency of those in tuple t_1.

Definition 5 (Data Timeliness Assessment Rule 2, DTAR2): Within relational database schema R, when considering tuples of the same entity encompassing associated or strongly correlated attributes within attribute sets X and Y, along with the derived rule ϕ from instance I in R where attribute X holds value X_ϕ and attribute Y holds value Y_ϕ, and for tuple t in I, with set X having attribute value X_t and set Y having attribute value Y_t, if both $X_t = X_\phi$ and $Y_t \neq Y_\phi$ hold, then Y_t is deemed outdated, thus failing timeliness requirements; otherwise, Y_t remains effective. This specific rule is articulated as:

$$\forall t \in R, \forall \phi_i \in \phi, (\exists X, Y \in R \wedge X_{t_i} = X_{\phi_j}) \rightarrow \begin{cases} Y_{t_i} \neq Y_{\phi_j}, currency(t_i) = 1 \\ Y_{t_i} = Y_{\phi_j}, currency(t_i) = 0 \end{cases} \quad \textbf{(6)}$$

3.3 Data Timeliness Evaluation

Assuming a relational data schema R comprises an instance I with D tuples, and if d tuples feature outdated attribute values or there are tuples with more current attribute values, the timeliness of instance I timeliness is assessed using the following formula. The value range lies within (0,1], where the closer to 1, the better timeliness performance:

$$Currency(I) = 1 - \frac{|d|}{|D|} \quad \textbf{(7)}$$

4 Attribute Value Dependency Rule Extraction and Data Timeliness Judgment Algorithm

In specific industry contexts, dependency relationships among data attributes can be derived from expert insights. However, delving into the level of data attribute values would substantially escalate manual efforts. Hence, this paper leverages machine learning techniques for extracting data attribute value dependency rules.

Algorithm 1. Algorithm for extracting conditional function dependency rules

For multi-factor industrial production time series data, the algorithm identifies each factor as an attribute, analyzes the dependency rules when the value of each attribute changes, and forms a set of attribute value dependency rules.

Input: A dataset R of strongly associated attributes with complete entity identification, consistency, and timestamp labels; representing the entity set E, a specific entity e, with $e \in E$ and an entity e corresponding to n records in the dataset [10]. Each record encompasses attribute sets X and Y, where here, X is denoted as the Left-hand Side (*LHS*), and Y as the Right-hand Side (*RHS*). Attribute sets X comprise h attribute values, and attribute sets Y consist of m attribute values.

Output: A set of attribute value dependency rules (AVD) containing r rules.

Algorithm:

```
AVD ← [];
for each e ∈ E do
   Re ← select * from R where eid = e.eid;
   for each i from 1 to n do
      ReLHS ← Re[i].LHS;
      ReRHS ← Re[i].RHS;
      LHS ←{}; RHS ← {};
      for each j from 1 to h do
         LHS AddDict ReLHS[j];
      end for;
      for each k from 1 to m do
         RHS AddDict ReRHS[k];
      end for;
      AVD AddDict {{'left':LHS},{'right':RHS}};
   end for;
   for i from 1 to r do
      r1 ← AVD[i];
      for j from 1 to r do
         r2 ← AVD[j];
         if r1 equals to r2 do
            delete r1;
         end if;
      end for
   end for;
return AVD;
```

Algorithm 2. Attribute value dependency rule update algorithm

Based on the set of attribute value dependency rules output from Algorithm 1, the attribute value dependency rule updating algorithm can extract attribute value dependency rules under different conditions according to different application scenarios, and finally generate a set of strongly associated attribute value dependency rules based on prioritized ordering and difference weights of the associated rules.

Input: Attribute value dependency rule library AVD, containing n data records, each linked to an attribute set *LHS*(X) with h attribute values, and an attribute set *RHS*(Y) with m attribute values. Additionally, new_AVD comprises newly extracted conditional function dependency rules from a fresh dataset, comprising n_1 data records, each tied to an attribute set *LHS*(X) with h attribute values, and an attribute set *RHS*(Y) with m attribute values.

Output: The updated attribute value dependency rule set updated_AVD.

Algorithm:

```
for each i in new_AVD from 1 to n1 do
   signal = 0;
   for each j in AVD from 1 to n do
```

```
        if(new_AVD[j][LHS] equals to AVD[i][LHS])&
        (AVD[j][RHS] NOT equals to new_AVD[i][RHS])
do
            AVD[j][RHS] ← new_AVD[i][RHS];
            signal ← 1;
            break;
        end if;
    end for;
    if signal equals to 0 do
        updated_AVD Add new_AVD[i];
end for;
updated_AVD Add AVD.
return updated_AVD;
```

Algorithm 3. Data timeliness judgment and evaluation

For each record in the strong correlation attribute value dependency rule set, data timeliness judgment and evaluation analyze the influence of each attribute value in the dependency relationship over time, and discards the unnecessary influence of the attribute value change based on the interaction's perturbation contribution analysis method, and ultimately comprehensively evaluates the validity of each attribute value dependency rule on the timeline.

Input: Attribute Value Dependency Rule Set (AVD) with n rules; data record set *data* for timeliness judgment containing m records, and each record encompassing data attributes present in AVD.

Output: Data record set *data* after completing data timeliness judgment, and the dataset's timeliness value *Currency* (*data*).

Algorithm:

```
initial currency; initial d;
for each i in data from 1 to m do
    for each j in AVD from 1 to n do
        if(data[i][LHS] equals to AVD[j][LHS]) &
        (data[i][RHS] NOT equals to AVD[j][RHS]) do
            data[i][currency] ← 1;
            d ← d+1;
        end if;
    end for;
    data[i][currency] ← 0;
end for;
Currency(data) = 1-(d/m);
return Currency(data), data;
```

5 Data Experiment and Result Analysis

5.1 Experimental Environment and Source Data

The experimental setup was configured on a single machine featuring a 16-core Central Processing Unit (CPU), 64GigaByte (GB) Radom Access Memory (RAM), and running on Community Enterprise Operating System 7.6 64-bit with Advanced Reduced Instruction-Set Computer Machine architecture. The system comprised a 300GB system disk and a 2x3.2TeraByte data disk, forming a 32-node large-scale data platform. The CPU configuration employed the Kunpeng 920 processor. Additionally, an Advanced Micro Devices, Inc. Ryzen 5 4600H 3.0Gigahertz processor, 16GB RAM, and Windows 10 64-bit operating system were employed on a personal computer. Testing datasets and rule sets were stored within a MySQL 5.7 database.

For the investigation into data timeliness assessment within industrial production, experimental source data was gathered from the thermal melting process of materials at a segment of Changhong Mold Plastics Co., Ltd. The data spanned from May 2022 to July 2022 and encompassed material, equipment, process, and quality data. Attributes such as "thermal melting material code, density/specific gravity, tensile strength, elongation, bending strength, volume resistivity, Rockwell hardness, Izod impact strength, melting point, heat resistance, voltage setting, current setting, temperature setting, component code, heating efficiency (initial efficiency, 500H efficiency, 1000H efficiency, 1500H efficiency, 2000H efficiency, 2500H efficiency, 3000H efficiency, 3500H efficiency, 4000H efficiency), nozzle temperature, drying temperature, screw speed, die temperature, molding temperature, molding shrinkage rate" constituted the data fields. A total of 417,638 records of injection molding material thermal melting process data for various material types were included.

Distinct material codes for thermal melting were employed as unique identifiers, associating production data with heating efficiency fluctuations over time, accounting for different materials, heating segments, and parameter configurations for various manufacturers. While data anomalies like "breakpoints" were observed due to factors like material quality or manual inspections, their presence was negligible within the extensive production dataset.

The current state of thermal melting in injection molding material production leans toward high automation and intelligence. During specific time frames, heating segment parameters (voltage, current, temperature), heating components, and duration remain relatively consistent for fixed thermal melting material types. However, as component usage time increases, process parameters fluctuate, or updates occur, heating efficiency evolves over time for diverse materials under distinct components. This evolution in component efficiency demonstrates clear temporal traits. Industries presently embrace automation and intelligence, including the industrial mold manufacturing sector. If manufacturers can establish a methodology to dynamically adjust thermal melting parameters, modify thermal melting materials, and manage component properties across equipment, materials, technology, and processes, sustained high thermal melting efficiency and consistent raw material shrinkage rates can be ensured. This operational stability contributes significantly to a company's competitive edge. Therefore, this paper employs the shrinkage rate of materials following heating under different conditions as a means to validate the proposed data timeliness assessment model.

Regarding data timeliness assessment within the scope of industrial production, the experimental source data originates from the factory workshop of Sichuan Changhong Mold Plastics Technology Co., Ltd. during a specific period in 2022.

5.2 Data Preparation

5.2.1 Data Cleaning

Upon initial statistical analysis of the source data, it became evident that certain components underwent prolonged and continuous heating, leading to material scrap or disruptions in the thermal melting process. As a result, instances of data anomalies or partial missing entries emerged, classifying this subset of data as noisy. Consequently, the remaining dataset comprised 348,029 records encompassing 23 categories of thermal melting materials and equipment data.

We collected a total of 918356 production data for processing and set up seven different data cleaning conditions and three missing data filling methods, including a tensor complementation model based on nonlocal similarity learning [16]. According to the hardware environment described in 5.1, after about one hour of processing, we obtained 673055 valid data.

5.2.2 Preliminary Data Processing

In the context of material thermal melting, higher melting points and molding temperatures correlate with broader temperature ranges sustainable within heating segments 1 and 2. Moreover, as the mold temperature approaches the critical temperature for material shaping, the molding shrinkage rate diminishes. To uncover optimal molding processes for thermal melting materials, the focus centers on heating efficiency of components and molding shrinkage across each component, heating segment, and mold structure. Furthermore, current heating patterns for thermal melting materials predominantly involve continuous heating within a single mold, lacking the flexibility to tailor thermal melting materials based on material attributes and component conditions. Hence, this experiment adopts the duration of continuous component heating as a statistical unit.

The objective of this experiment is to assess the timeliness of process parameters and heating efficiency linked to material-specific component heating. Additionally, the shrinkage rate subsequent to thermal melting material molding is primarily influenced by heating temperature, drying temperature, screw speed, and mold temperature. Consequently, the requisite data fields encompass “material type code, component code, heating segment 1 (voltage, current, temperature), screw speed, mold temperature, heating duration, heating efficiency, shrinkage rate”. The data sample is presented in Table 1 below.

Table 1. Hot melt data of injection molding production materials

Material	Component	Section1			Heating						Mold temperature	Screw rotating	Standard shrinkage
		U	I	Temp	Time	Efficiency	Time	Efficiency	Time	Efficiency			
ABS	H	220	10	180	600	0.91	1000	0.86	1400	0.81	55	40	0.54
ABS	S	220	12	180	600	0.93	1000	0.87	1400	0.82	70	45	0.66
ABS	M	220	15	180	600	0.89	1000	0.83	1400	0.77	80	60	0.61
PS	H	220	10	200	600	0.91	1000	0.85	1400	0.79	60	50	0.42
PS	S	220	12	200	600	0.93	1000	0.88	1400	0.83	30	40	0.57
PS	M	220	15	200	600	0.90	1000	0.84	1400	0.77	55	50	0.41
PP	H	220	10	180	600	0.91	1000	0.86	1400	0.81	50	55	1.29
PP	S	220	12	180	600	0.93	1000	0.87	1400	0.82	70	65	1.81
PP	M	220	15	180	600	0.89	1000	0.83	1400	0.77	80	85	2.40
PC	H	220	10	260	600	0.91	1000	0.86	1400	0.81	80	30	0.92
PC	S	220	12	260	600	0.93	1000	0.87	1400	0.82	90	40	0.85
PC	M	220	15	260	600	0.89	1000	0.83	1400	0.77	100	50	0.84
PVC	H	220	10	160	600	0.91	1000	0.86	1400	0.81	25	30	0.2
PVC	S	220	12	160	600	0.93	1000	0.87	1400	0.82	40	30	0.5
PVC	M	220	15	160	600	0.89	1000	0.83	1400	0.77	55	30	0.5

5.2.3 Determination of Thermal Melting Process Parameter Efficiency Thresholds for Materials

Drawing from the widely acknowledged Pareto principle in production and daily contexts, this paper establishes a ranking order for thermal melting efficiency among diverse materials, components, process parameters, and thermal melting durations. The arrangement proceeds from high to low efficiency. Consequently, parameters such as heating segment voltage, current, temperature, thermal melting duration, and mold temperature linked with thermal melting segments exhibiting consistent efficiency above 0.9 and shrinkage rates within regulatory boundaries are classified as “Optimal” parameter configurations for the material’s thermal melting effect. Among the remaining thermal melting configurations, those demonstrating efficiency spanning 0.7 to 0.9 continuously and shrinkage rates within regulatory boundaries are categorized as “Good.” Conversely, configurations featuring efficiency below 0.7 or materials displaying shrinkage rates exceeding regulatory specifications are denoted as "Poor."

After conducting statistical analysis of the data, the thresholds for thermal melting efficiency process parameters are presented in Table 2 below. Here, "heating config" signifies the "thermal melting configuration," "efficiency_good" denotes the threshold separating "Medium" and "High" thermal melting efficiencies, and "efficiency_medium" defines the threshold distinguishing "Low" and "Medium" thermal melting efficiencies.

Table 2. Efficiency threshold of process parameters for material hot melting

heating_config (type)	efficiency_ _good	efficiency_ _medium
ABS-H	(220,18, 210,800)	(220,10, 200,2600)
ABS-S	(220,20, 220,1000)	(220,10, 180,3000)
ABS-M	(220,15, 220,600)	(220,15, 200,2000)
PS-0	(220,10, 200,800)	(220,25, 260,2200)
PS-P	(220,10, 180,400)	(220,12, 210,1600)
PS-S	(220,15, 190,1000)	(220,25, 220,3000)
PP-H	(220,22, 250,800)	(220,10, 260,2400)
PC-O	(220,15, 300,800)	(220,12, 280,2400)
PC-S	(220,18, 320,800)	(220,12, 280,3000)
PC-M	(220,10, 260,600)	(220,25, 360,1800)
PVC-M	(220,18, 175,600)	(220,20, 180,2000)
PVC-O	(220,15, 170,800)	(220,15, 170,2400)
PVC-P	(220,12, 165,400)	(220,15, 170,1800)

5.2.4 Dataset Partitioning

To fulfill the experimental requirements, this study extracts 23 categories of materials from the thermal melting material dataset and divides them into three distinct datasets. Dataset 1 encompasses 86,619 instances of material thermal melting data spanning from May 1st to May 31st, 2022. Dataset 2 comprises 193,566 instances of material thermal melting data covering June 1st to June 30th, 2022. Dataset 3 contains 67,844 instances of material thermal melting data recorded from July 1st to July 31st, 2022. For Dataset 1, 75% of the instances are randomly chosen for extracting attribute value dependency rules, while the remaining 25% are retained for time-effectiveness assessment. In Dataset 2, instances corresponding to subjects extracted from Dataset 1 are selected for validating data timeliness. The remaining material thermal melting data in Dataset 2 are used for extracting new material thermal melting process attribute value dependency rules. Lastly, Dataset 3 is utilized to validate the accuracy of time-effectiveness evaluation in this experiment.

5.3 Performance Testing of Data Attribute Value Dependency Rule Extraction Algorithm

5.3.1 Testing the Size of Extracted Rule Sets

The performance evaluation of the data attribute value dependency rule extraction algorithm involves dividing the dataset into ten sets of material thermal melting data. Each set undergoes ten rounds of algorithm testing. These ten material thermal melting datasets are randomly drawn from Datasets 1 and 2. The size of the datasets increases progressively according to a specific gradient. The smallest dataset contains thermal melting data for five material types, while the largest encompasses thermal melting data for 23 commonly used materials. Despite datasets of the same size, variations in the sizes of the extracted rule sets might arise due to the random nature of thermal melting material dataset extraction. In this experiment, the smallest extracted rule set contains 18 rules, while the largest comprises 1,299 rules. This relationship is visually depicted in Figure 1.

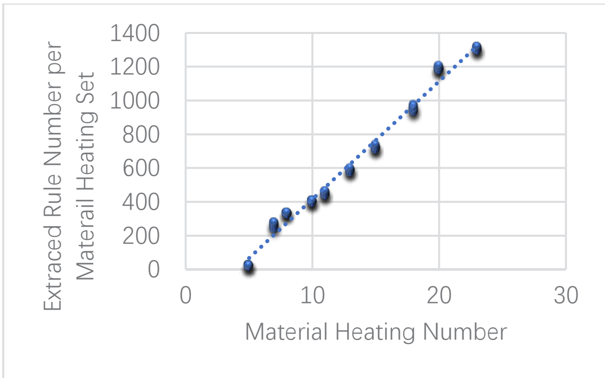

Figure 1. Rule number extracted from per material heating set

From the figure, it can be seen that the size of the rule set extracted from the material hot melt dataset is basically proportional to the size of the material hot melt dataset, and the number of rules increases with the increase of the material hot melt quantity.

5.3.2 Testing the Efficiency of Data Attribute Value Dependency Rule Extraction Algorithm

The results of 100 tests on the data attribute value dependency rule extraction algorithm are presented in the subsequent figures. The horizontal axis represents the number of extracted rules, while the vertical axis illustrates the time expended for rule extraction. Figure 2 showcases the correlation between the number of extracted rule sets and the time consumed for extraction, while Figure 3 portrays the relationship between the number of extracted rule sets and the time taken for individual rule extraction. Under optimal conditions, the algorithm can extract 12,562 rules per millisecond, whereas under less favorable conditions, only 122 rules per millisecond can be extracted.

Due to limitations in computer storage and computational resources, the time consumption for rule extraction does not exhibit an absolute linear growth trend. The trend line included in the scatter plots confirms the direct relationship between the number of rules and the total time taken. Furthermore, the average time cost per rule diminishes as the number of rules increases. After reaching a certain threshold, the cost per rule extraction remains constant, validating the association between the number of rules and the total time taken.

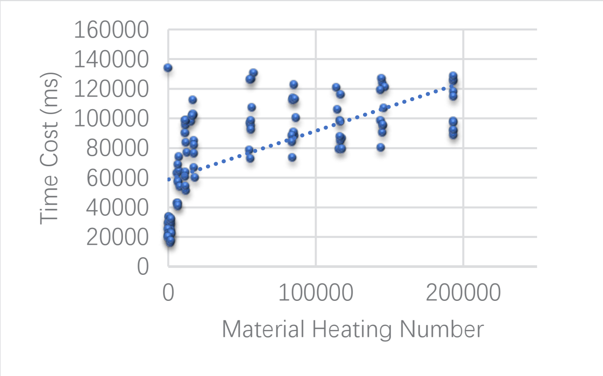

Figure 2. Time cost for extracted rules extraction

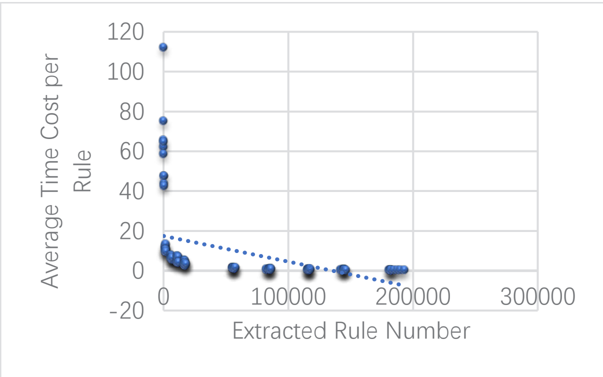

Figure 3. Time cost for per rule

5.4 Timeliness Judgment for Injection Molding Material Thermal Melting Process Parameters

5.4.1 Extraction of Injection Material Heating Efficiency CAVD Rules

As per the dataset partitioning strategy delineated earlier, 75% of the material thermal melting process data (comprising five material types from the main material group A) are randomly selected from Dataset 1. From this subset, heating efficiency CAVD rule set 1 is extracted. The left-hand side of the rules relies on dependent attributes such as "Material Thermal Melting Configuration" and "Heating Duration", while the right-hand side is dependent on the depending attribute "Heating Efficiency (High/Medium/Low)". The CAVD rule extraction algorithm has been detailed in Section 4.1. This extraction process yields 64,277 rules with high heating efficiency, 15,363 rules with medium heating efficiency, and 6,979 rules with low heating efficiency, as outlined in Table 3.

Table 3. Examples of CAVD rules

C (Type)	X (Time)	Y (Efficiency)
ABS-H	358.6	high
ABS-S	1858.3	medium
ABS-M	366.7	high
PS-H	1362.7	medium
PS-S	2832.7	low
PS-M	600.0	High
PP-H	1487.3	medium
PP-S	3145	low
PP-M	0.0	high
PC-H	2290.9	low

5.4.2 Labeling of Injection Material Thermal Melting Configurations

Utilizing the extracted CAVD rules from the previous section, the remaining 25% of the material thermal melting process data (pertaining to the mainstream material group B, encompassing eight distinct materials not covered in Section 5.4.1) are labeled based on heating efficiency. If the heating efficiency, heating parts, parameter configurations, and material-part heating efficiency in the CAVD rule for a material are identical, and the heating duration and CAVD rule heating duration match, then the thermal melting process configuration for that material is the same as the process parameters in the CAVD rules. Labeled thermal melting process configuration data tuples are illustrated in Table 4.

Table 4. Examples of material hot melt process configuration type

Material	Type	Time	Label (Efficiency)
AS	AS-H	401.8	high
HIPS475	HIPS475-S	3079.2	low
LDPE	LEPE-M	644.4	high
HDPE	HDPE-H	1573.3	medium
PMMA	PMMA-S	2417.4	low
PA6	PA6-M	377.1	high
PA66	PA66-H	1208.6	medium
POM	POM-S	3107.0	low

5.4.3 Extraction and Consolidation of New CAVD Rules

Adhering to previous dataset partitioning, injection production thermal melting process data for the main material group A are extracted from Dataset 2 to derive new heating efficiency CAVD rules. This process yields 67,582 new rules, which are combined with rule set 1 to constitute rule set 2.

5.4.4 Timeliness Judgment and Verification of Injection Production Material Thermal Melting Configurations

Over time, material thermal melting configurations may undergo changes. In response, the heating efficiency and shrinkage rates of material group B are examined based on the rules in rule set 2. Any changes indicate outdated thermal melting configurations, while no changes indicate the continued effectiveness of high-efficiency, high-quality material thermal melting configurations. The timeliness judgment algorithm elucidated in Section 4.3

is applied. As per the algorithm, around 16% of material group B (material group C) possesses outdated material thermal melting configurations, while 82.87% of material group B (material group D) maintains effective thermal melting configurations. The timeliness evaluation value is 0.8287.

In Dataset 3, material thermal melting configurations, heating efficiency, and material shrinkage rates are statistically analyzed. In material group C, 93.46% of materials manifest significant changes in thermal melting efficiency and shrinkage rates. In material group D, 98.11% of materials exhibit minimal changes in thermal melting efficiency and shrinkage rates. This underscores the effectiveness of the proposed timeliness judgment model in assessing the timeliness of attribute-related data sets.

6 Conclusion and Future Research

The assessment of data timeliness holds paramount significance in the realm of data quality evaluation, as utilizing outdated data can result in significant organizational losses. This paper has introduced a novel approach for timeliness judgment grounded in attribute value dependencies, enabling the determination of data validity within expected temporal boundaries. The proposed approach is rooted in the extraction of attribute value dependency rules from interconnected attributes within the dataset. It extrapolates temporal changes in attribute values via the temporal shifts in these dependency rules. The method furnishes a timeliness judgment rule and an evaluation model for data timeliness. The conducted experiments corroborate the precision of the timeliness judgment rule and the efficacy of the evaluation model. The research results include extracting conditional function dependency rules and data timeliness evaluation, and validating the effectiveness of the algorithms by industrial production data. The conditional function dependency rules are used to extract the influence factors and the dependency rules, helping to determine the multiple factors that affect industrial production efficiency and quality. The data timeliness evaluation can be used to filter out effective data from massive industrial production data, and the core influencing factors with better time-efficient to make the massive industrial production data easy to use. Eventually, the research results can effectively improve the use value of industrial data in the entire manufacturing process, and drive manufacturing towards intelligence.

Moving forward, future research endeavors within this field will delve into the following areas: (1) Formulating timeliness judgment rules to accommodate data tuples featuring multiple sets of interrelated attributes; (2) Enhancing the efficiency of learning algorithms for attribute value dependency rules and exploring more efficacious sampling techniques to curtail the scale of rule learning training sets; (3) Tackling attribute value rectification issues concerning attributes beyond those with established correlations within data tuples; (4) Design and implementation of a novel privacy-preserving tensor decomposition method on secure encryption of big data attribute values [17].

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant No. 2020YFB1707900 and 2020YFB1711800; the National Natural Science Foundation of China under Grant No. 62262074, U2268204 and 62172061; the Science and Technology Project of Sichuan Province under Grant No. 2022YFG0159, 2023ZHCG0014, 2023ZHCG0011, 2022YFG0155, 2022YFG0157.

References

[1] P. Oliveira, F. Rodrigues, P. R. Henriques, A Formal Definition of Data Quality Problems, *MIT International Conference on Information Quality*, Cambridge, MA, 2005, pp. 1-14.

[2] X. O. Ding, H. Z. Wang, X. Y. Zhang, J. Z. Li, H. Gao, Association Relationships Study of Multi-dimensional Data Quality, *Journal of Software*, Vol. 27, No. 7, pp. 1626-1644, March, 2016.

[3] C. Cappiello, C. Francalanci, B. Pernici, A Model of Data Currency in Multi-Channel Financial Architectures, *MIT International Conference on Information Quality*, Cambridge, MA, USA, 2002, pp. 106-118.

[4] M. H. Li, J. Z. Li, S. Y. Cheng, Y. B. Sun, Uncertain Rule based Method for Determining Data Currency, *IEICE Transactions on Information and Systems*, Vol. E101-D, No. 10, pp. 2447-2457, October, 2018.

[5] Y. Liang, X. L. Duan, Y. J. Ding, X. F. Kou, J. C. Huang, Data Mining of Students' Course Selection based on Currency Rules and Decision Tree, *Big Data and Computing*, Guangzhou, China, 2019, pp. 247-252.

[6] W. F. Fan, F. Geerts, J. Wijsen, Determining the Currency of Data, *ACM Transactions on Database Systems*, Vol. 37, No. 4, pp. 1-46, December, 2012.

[7] M. H. Li, J. Z. Li, Algorithms for Improving Data Currency, *Journal of Computer Research and Development*, Vol. 52, No. 9, pp. 1992-2001, September, 2015.

[8] W. F. Fan, F. Geerts, N. Tang, W. Y. Yu, Conflict Resolution with Data Currency and Consistency, *Journal of Data and Information Quality*, Vol. 5, No. 1-2, pp. 1-37, September, 2014.

[9] M. H. Li, J. Z. Li, A Minimized-rule based Approach for Improving Data Currency, *Journal of Combinatorial Optimization*, Vol. 32, No. 3, pp. 812-841, October, 2016.

[10] X. L. Duan, B. Guo, Y. Shen, X. Q. Dong, H. Zhang, Data Repair Algorithm Based on Currency Rules, *Journal of Software*, Vol. 30, No. 3, pp. 589-603, March, 2019.

[11] X. O. Ding, H. Z. Wang, Y. T. Gao, J. Z. Li, H. Gao, Determining the Currency of Dynamic Data, *ACM Turing Celebration Conference*, Shanghai, China, 2017, pp. 1-6.

[12] X. O. Ding, H. Z. Wang, Y. T. Gao, J. Z. Li, H. Gao, Efficient Currency Determination Algorithms for Dynamic Data, *Tsinghua Science and Technology*, Vol. 22, No. 3, pp. 227-242, June, 2017.

[13] X. G. Liu, L. S. Wu, C. Dai, H. C. Chao, Compressing CNNs Using Multilevel Filter Pruning for the Edge Nodes of Multimedia Internet of Things, *IEEE Internet of Things Journal*, Vol. 8, No. 14, pp. 11041-11051, July, 2021.

[14] R. Huo, H. Z. Wang, R. Zhu, J. Z. Li, H. Gao, Map-reduce based Entity Identification in Big Data, *Journal of Computer Research and Development*, Vol. 50, No. S2, pp. 170-179, November, 2013.

[15] H. Z. Wang, W. F. Fan, Object Identification on Complex Data: A Survey, *Chinese Journal of Computers*, Vol. 34, No. 10, pp. 1843-1852, October, 2011.

[16] C. Dai, Y. Zhang, Z. G. Zheng, A Nonlocal Similarity Learning-based Tensor Completion Model with Its Application in Intelligent Transportation System, *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-12, September, 2023. https://ieeexplore.ieee.org/document/10247151

[17] J. Feng, L. T. Yang, Q. Zhu, K. K. R. Choo, Privacy-Preserving Tensor Decomposition over Encrypted Data in a Federated Cloud Environment, *IEEE Transactions on Dependable and Secure Computing*, Vol. 17, No. 4, pp. 857-868, July-August, 2020.

Biographies

Jun Tang is studying for a PhD at the College of Computer Science, Sichuan University. He is the head of Changhong's big data team. His main research areas are big data modelling and data governance.

Bing Guo is a professor and doctoral supervisor at the College of Computer Science, Sichuan University. He earned his PhD degree from University of Electronic Science and Technology of China in 2002. His main research interests are personal big data management and blockchain technologies.

Yan Shen is a professor and doctoral supervisor at the School of Control Engineering, Chengdu University of Information Technology. She obtained her PhD degree from University of Electronic Science and Technology of China in 2004. Her research primarily focuses on intelligent terminals and instruments.

Cheng Dai is an associate professor in the school of Computer Science, Sichuan University. He received the Ph.D. degree from University of Electronic Science and Technology of China, in Jun. 2021. His current research interests include deep learning, edge intelligence and tensor decomposition.

Yuming Jiang is a professor in the College of Computer Science, Sichuan University in China. He received his MS and Ph.D. degrees in mechanical manufacture and automation from the Sichuan University in 1990 and 2001, respectively. His current research interests include intelligent manufacturing, big data and block chain.