Journal of Computer and Communications, 2014, 2, 17-24
Published Online December 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.214002

Scientific Research

A Scheme for Mining State Association Rules of Process Object Based on Big Data

Qiaoyun Song[1], Qingbei Guo[1], Kai Wang[1], Tao Du[1], Shouning Qu[2*], Yong Zhang[3]

[1]School of Information Science and Engineering, University of Jinan, Jinan, China
[2]Information Network Center, University of Jinan, Jinan, China
[3]School of Electrical Engineering, University of Jinan, Jinan, China
Email: songqiaoyun.1223@163.com, [*]qsn@ujn.edu.cn

Received 17 October 2014; revised 12 November 2014; accepted 26 November 2014

Abstract

This paper devises a scheme which can discover the state association rules of process object. The scheme aims to dig the hidden close relationships of different links in process object. We adopt a method based on difference and extremum to compute the timing. Clustering is used to classifying the adjusted data, and the next is associating the clusters. Based on the rules of clusters, we produce the rules of links. Association degrees between each two links can be determined. It is easy to get association chains according to the degree. The state association rules that can be obtained in accordance with association rules are the final results. Some industry guidance can be directly summarized from the state association rules, and we can apply the guidance to improve the efficiency of production and operational in allied industries.

Keywords

Process Object, Timing, Association Chain, State Association Rule

1. Introduction

Big data has 4 characteristics [1] which can be summarized as "4V": volume, variety, velocity and value. Moreover, big data may not be stored in fixed database; it also spread out in network space in different places. Semi-structured data or unstructured data is the main type of big data, so big data is complexity. Undoubtedly, all these characteristics bring more difficulty to storage, calculation and knowledge discovery. The process industry plays a dominant role in country economy. It produces real-time dynamic data and accumulates large amount of historical data. The data is an important component of big data. We can extract knowledge and interesting infor-

[*]Corresponding author.

How to cite this paper: Song, Q.Y., Guo, Q.B., Wang, K., Du, T., Qu, S.N. and Zhang, Y. (2014) A Scheme for Mining State Association Rules of Process Object Based on Big Data. *Journal of Computer and Communications*, **2**, 17-24.
http://dx.doi.org/10.4236/jcc.2014.214002

mation from big data while the knowledge is hidden, unknown, but potentially useful.

In process industry, industrial installation is composed of multiple operation unites or equipment generally. The input of downstream unite is usually output of upstream unite. To make full use of the equipment capacity and mining enterprise production potential, process industry should ensure failure-free operation of the equipment. However, researchers intuitively obtain the correlation of the data simply through regular analysis. No effective algorithm is adopted to discover hidden knowledge, so we get less regulation or rule from big data.

2. Related Work

The research on association rules is paid more and more attention by many researchers. Association rule mining was first introduced by reference [2] [3]. Since then, it has been extensively studied.

A method proposed by reference [4], which can find concept from time series, could be the start of the time series data mining. Firstly, it used the property of dynamical system behind time series data to delay the time series. Then they clustered the result, and applied the clusters to machine learning. But the algorithm did less work in the field of association rules. A novel divide and conquered two-phase algorithm is presented in [5]. It guaranteed to find all good rules efficiently. The paper also proposed an optimization technique that drastically improved the speed, and discussed how to maintain the rules. Reference [6] developed an algorithm that partitions the domain of items according to their correlations. It described a mining algorithm that carefully combined partitions to improve the efficiency of the mining procedure. The authors raised a chain structure to store frequent item sets in [7]. The algorithm can promotes the efficiency of mining frequent pattern. Reference [8] studied the question of incremental updating for mining association rules in large transaction database. At the same time, the authors presented an incremental updating algorithm based on frequent pattern tree to deal with the update of association rule after the change. In [9], authors put forward a parallel FP-Growth algorithm based on composite list mining under the cloud environment. The algorithm use cloud computing to handle the big data, and mine frequent patterns based on composite list to instead of constructing FP-growth tree or condition FP-growth tree. These algorithms can just get rules like "$a \to b$", but they are inability to the relationship of links in process object.

There are many problems involved in discovering hidden knowledge, such as computing the timing of the process object, the classifying of the data, the producing and using of the rule and so on. Against these problems, this paper proposes a scheme. The scheme adopts various data mining algorithms and technologies to discover the state association rules of process object based on association chains. From the state association rules we can intuitively know how a state change of a link influences the others. According to these rules, people can give the process industries professional guidance in fault analysis, failure detection, optimal state estimation and so on.

3. Definitions of Process Object

For convenience of the following analysis, this paper gives several definitions.

Definition 1. An object composed of n links is called process object, which the sample data of all links can form a time series. Assume that we have a process object χ, if $T_M = \{t_1, t_2, \cdots, t_m\}$ and $t_1 < t_2 < \cdots < t_m$, then

$$\chi = \left\{ X_1\left(x_1(t_1), x_1(t_2), \cdots, x_1(t_m)\right), X_2\left(x_2(t_1), x_2(t_2), \cdots, x_2(t_m)\right), \cdots, X_n\left(x_n(t_1), x_n(t_2), \cdots, x_n(t_m)\right)\right\} \quad (1)$$

where $X_i, X_j\ (i, j = 1, 2, \cdots, m)$ are the sample data of any two links in $T = \{t_1, t_2, \cdots, t_m\}$, and there exists $f_{ij}(x)$ that makes

$$X_i(t) = f_{ij}\left(X_j\left(t + \Delta t_{ij}\right)\right), t \in T \quad (2)$$

Δt_{ij} is the response time of X_j when X_i has a change.

Definition 2. A unidirectional chain which is composed of different links based on correlation degree is called association chain.

Definition 3. A rule likes a chain which element is the state of X_i is called state association rule.

Assume that process object χ include n links. The data of all links is sampled in $T = \{t_1, t_2, \cdots, t_m\}$, where $t_1 < t_2 < \cdots < t_m$. The following analysis of this paper is based on the assumption, and undoubtedly, the assumption is reasonable.

4. Scheme Design

This paper devises a scheme to find the implicit state association rule of process object. The scheme consists of five main steps including data sampling, timing analysis, clustering, association rule mining, association chain mining and state association rule generation. In timing analysis step, a novel method based on counting was proposed to determine time series and time delay of different links. In clustering step, data collected at the same time was divided into k classes by k-means clustering algorithm. The novel step uses silhouette coefficient based on cohesion degree and separation degree as the clustering criteria. In association chain step, cluster set was organized into the association chain containing only a single chain or the association tree containing multiple chains. Using these association chains, state association rules are easily obtained. The state association rule reflects the relationship of different links.

The scheme is shown in **Figure 1**.

4.1. Sampling

In this article, difference serves as sampling criteria and reflects data changes over time. Obviously, the larger the variation of data, the more rich the information contained in these data. In our practical application, original data are divided equally into m segments. $\Delta\chi$ indicates the variation of χ, $\Delta\chi$ is defined as the sum of absolute first-order difference. The segment with the largest $\Delta\chi$ was selected. The selected segment is noted as χ_M, and the period is noted as T_M. Compared with other segments, χ_M contains the most information, so χ_M can represent the raw data.

4.2. Timing

An idea on the basis of difference and extremum is put forward in this section to calculate the timing. In the meantime, the delay time between different links can be known. In process industry, the change of any link will influence the others. Imagining one link has great fluctuation, it must cause some changes in other links. That is to say, there must have the corresponding extremums turned up in some links while a extremum appeared in one

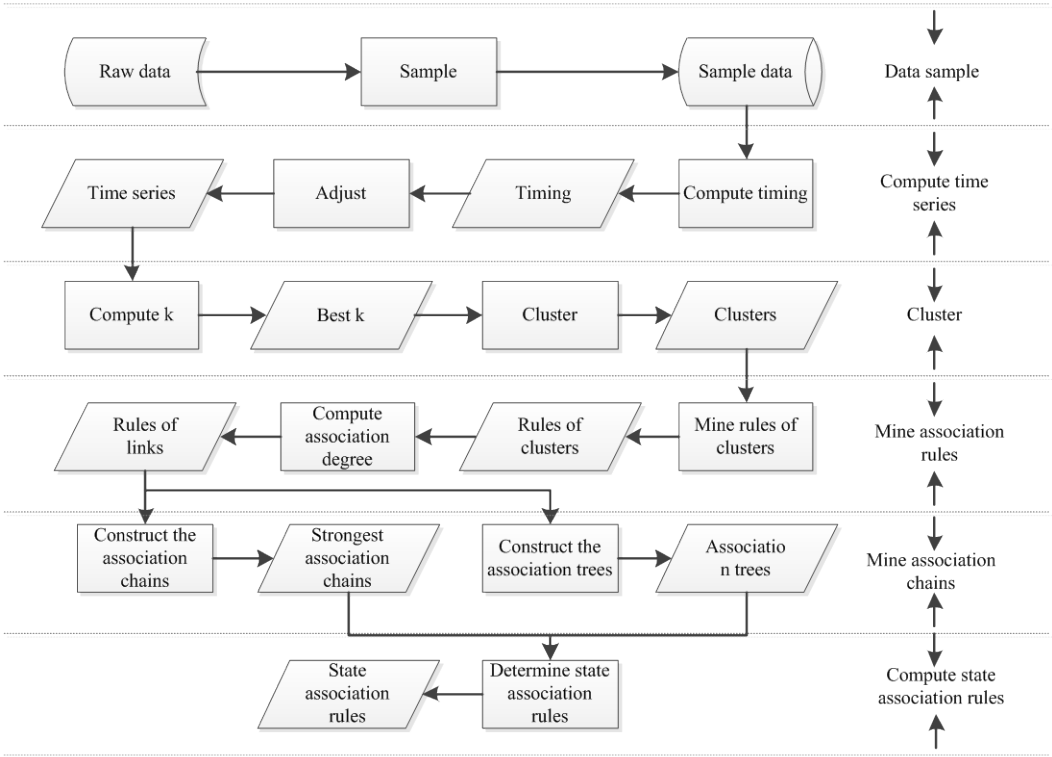

Figure 1. The flow chart.

link. The interval between different extremums is the delay time. Let Δt_{ij} be the delay time between X_i and X_j. In process industry, the speed of passing information between adjacent links is fast, and almost the delay time between any two links is shorter than the interval between different extremums. So we can calculate Δt_{ij} according to the difference of extremums. In practical case, fluctuation can spread rapidly and also the sample data exist a lot of noise. In order to reduce error and increase measurement precision, the delay time Δt_{ij} which makes them the most frequent can be treated as actual delay time.

By now, the time series data is emerged. And then we can adjust the data based on the delay time. Assume the order of all links after adjusted is

$$L = X_1' \to X_2' \to \cdots \to X_m' \tag{3}$$

L is unidirectional.

4.3. Clustering

In clustering step, *k*-means algorithm was adopted. After *k*-means, each link is separated to different classes with their best *k*. Each class represents a state of the link. So, every link can be simplified as *k* states. The biggest benefit would be to reduce the amount of computation, thereby increase the practicability of this method.

To determine *k*, silhouette coefficient [3] denoted as s_k based on cohesion degree and separation degree is adopted. For any one clustering result with *k* classes, the silhouette coefficient is

$$s_k = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max\left(a_i - b_i\right)} \tag{4}$$

where a_i is the average distance of *i*-th sample point and other sample points in the same class, b_i is the average distance of *i*-th sample point and other centers of class. An ideal clustering result should have the minimum cohesion degree and the maximum separation degree. There will be a low computational because of the large amount of data. For this reason, one link with the largest variation can be selected to determine *k*.

Suppose the best *k* of link $X_i\left(i=1,2,\cdots,m\right)$ is denoted as $k_i\left(i=1,2,\cdots,m\right)$. After *k*-means, all links are separated to $k_i\left(i=1,2,\cdots,m\right)$ classes. For the links, the $k_i\left(i=1,2,\cdots,m\right)$ classes can represent all the states of corresponding links.

4.4. Association Chain

Apriori algorithm is a most valuable frequent item sets data mining algorithm to find boolean association [2]. Apriori algorithm can only find one-dimensional boolean association rules while the state association rule of process object is multidimensional rules. Interdimension association rule mining algorithm based on Apriori is adopted. If the number of element of frequent predicate set is greater than 2, then the antecedent or consequent of corresponding association rule has multiple predicate. In that case, this rule cannot be expressed as one-to-one correspondence relations. This paper proposes an idea to set the number of element of frequent predicate set be 2. Then the association rules of the links' states are obtained. According to the interestingness and support of the rules of clusters we can gain a value which can be the correlation between two links, and also the rules of links obtained.

At present, we have already known the rules between links and the association degree. Based on this information, choose the rules to structure association rule which have the biggest association degree and satisfy the order L. We call the association rule as the strongest association rule that express a relationship between different links. Put any one of the links as the first link of the chain, and it will construct a chain. There will be $n-1$ strongest association chains with which head nodes are replaced by

$$\left\{X_i \middle| i=1,2,\cdots,n\right\} \cap \left\{X_i \neq X_m'\right\} \tag{5}$$

In order to fully exploit the possible hidden relationships of all links, a binary tree, which we called association tree, need to be constructed. The tree will be constructed on the basis of association chains which can show the relationship of different links. All links can be included in the tree, and every branch of the tree is an association chain.

Suppose any one of association rules is denoted by φ_i. Assuming that φ_i includes d links, it can be represented as

$$\varphi_i = X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_d \tag{6}$$

4.5. State Association Rule

From the association rules we can know that there exist mutual influences and relations between links, but it is unable to determine how a link state influence the others. In view of the problem, this section provides an idea based on difference to determine the relationship between adjacent links on association rules. Generally, the state of numeric data can be distributed into 3 types: rise, fall and unchanged. $\Delta\chi$ can express the state. According to $\Delta\chi$, we can get the state of all the association rules, that is to say the state association rules like chains obtained. Count the number of state association rule and the number can represent the probability of the rule's state.

The state value of any one of φ_i at t_j is

$$\varphi_i(t_j) = \{\zeta_1(t_j), \zeta_2(t_j), \cdots, \zeta_d(t_j)\} \tag{7}$$

where $\zeta_i(t_j)$ is the state value of X_i at t_j. Count all state value of φ_i in T_M, and then get all the state association rules. Suppose $\psi_i(j)$ represents any one of state association rule φ_i. If φ_i generate h rules in total, we can know

$$\psi_i = \{\psi_i(1), \psi_i(2), \cdots, \psi_i(h)\} \tag{8}$$

The probability of $\psi_i(j), (j = 1, 2, \cdots, h)$ is

$$p(\psi_i(j)) = \frac{N(\psi_i(j))}{N(\psi)} \tag{9}$$

where $N(\psi_i(j))$ is the number the state $\psi_i(j)$ occurrences in T_M, and $N(\psi)$ is the total number of transactions.

The number from big to small is the process that the state of object from normal to abnormal. We can directly gain some industry guidance from the state association rules, and then give some guidance to improve the efficiency of production and operational in allied industries.

5. Experiment Result and Analysis

We have performed some experiments to make sure that our method works effective. Power generation system of the electric power is a typical process industry system. The whole process flows of power system is a process object. The historical data of a subsystem of a power plant are selected to be the experimental data. 789 days of data are filtered down to 1,070,008 pieces of data. The time interval of data acquisition is 1 min. There are 8 links with their names list in **Table 1**, denoted as $\{X_1, X_2, \cdots, X_8\}$. We have mined the association chains from the data, and have computed the state association rules.

5.1. Experiment Result

The rough industry process of all links is gained after timing. The process is shown as following.

$$L = \{X_1, X_2, X_4, X_6, X_8\} \rightarrow \{X_7\} \rightarrow \{X_3\} \rightarrow \{X_5\} \tag{10}$$

According to L, adjusted the sample data to ensure that the data of all links in the same time is sequential. Clustered all the links into 1 - 10 classes. Their best k is determined by silhouette coefficient with the 10 clustering results. Based on the best k to separate $X_i (i = 1, 2, \cdots, 8)$ into $k_i (i = 1, 2, \cdots, 8)$ classes by *k*-means algorithm. Mined the association rules with one antecedent and one consequent by interdimension association rule mining algorithm based on Apriori. From this step, we obtained the rules between clusters and their support and interestingness. Computed the association degree between two links with the support and interest. From the point of this result, we gained the strongest association chains like **Table 2**.

The association tree contains more possibility of the relationships of all links, such as the association tree of X_1 (**Figure 2**).

Table 1. The name of all links.

Links	Name
X_1	10HNA10CQ1013S
X_2	10HNA10CQ1013S
X_3	01_q2
X_4	01_q4
X_5	01_q3
X_6	01_Qnetar
X_7	01_Vdaf
X_8	MSTMFLOW

Table 2. The strongest association chains.

Links	Association Chains
X_1	$X_1 \to X_8 \to X_5$
X_2	$X_2 \to X_1 \to X_8 \to X_5$
X_3	$X_3 \to X_4 \to X_5$
X_4	$X_4 \to X_5$
X_5	--
X_6	$X_6 \to X_4 \to X_5$
X_7	$X_7 \to X_4 \to X_5$
X_8	$X_8 \to X_1 \to X_7 \to X_4 \to X_5$

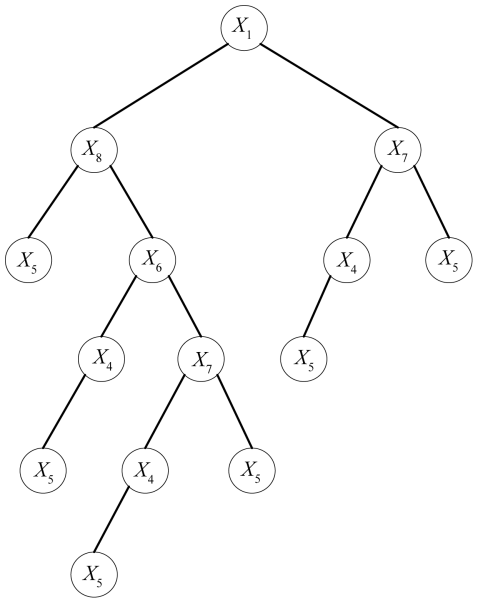

Figure 2. The association tree of X_1.

5.2. The Analysis of Experiment Results

Take the strongest association chain begins with X_7 for example, the association chain can have 27 states in theory. The state association rules of the chain we computed is shown in **Table 3**.

As we all know, there are only a few states in a process industry. From **Table 2** we can know that there are only 19 states occurred which just proves the phenomenon. The percentage shows the probability of all state as-

Table 3. The state association rules of $\{X_7 \rightarrow X_4 \rightarrow X_5\}$.

Number	X_7	X_4	X_5	Percentage	State
1	unchanged	fall	fall	20.876	normal
2	unchanged	rise	rise	20.467	normal
3	unchanged	fall	rise	17.442	transition
4	unchanged	rise	fall	17.109	transition
5	unchanged	unchanged	fall	11.17	transition
6	unchanged	unchanged	rise	7.392	transition
7	rise	rise	fall	1.271	abnormal
8	fall	fall	fall	0.758	abnormal
9	fall	fall	rise	0.741	abnormal
10	rise	fall	fall	0.55	abnormal
11	rise	unchanged	rise	0.399	abnormal
12	rise	rise	rise	0.389	abnormal
13	unchanged	unchanged	unchanged	0.354	abnormal
14	fall	rise	rise	0.352	abnormal
15	fall	rise	fall	0.307	abnormal
16	rise	fall	rise	0.238	error
17	unchanged	rise	unchanged	0.183	error
18	fall	unchanged	fall	0.001	error
19	unchanged	fall	unchanged	0.001	error

sociation rules.

The first two state association rules with the bigger percentage are considered to be rules with the normal states in the process. Both the two have the common feature that X_4 and X_5 have the same state while X_7 remain unchanged. The results tells that X_7 is likely to be a property and it has a close relationship with both X_4 and X_5. X_4 and X_5 are positive since the variation of X_4 can lead to similar variation. Actually, X_7 is exactly a percentage of coal, and X_4 and X_5 have a close relationship. The total percentage of rules from 7 to 15 is less than 6%. All those rules are thought to be abnormal. It must be caused by some faults for that X_7 has different changes. The rules of 3 to 6 are considered to be the transition states. Because of the situation that X_4 and X_5 have the different state while X_7 still keep the unchanged state, we infer those rules in the transition state from normal to abnormal. The last 4 rules account for less than 1% can be ignored. It can be regard as the error data caused by measurement or other reasons.

Each branch of association tree is an association chain. The results of association trees are likely to the above analysis.

6. Conclusions and Outlook

In this paper, we proposed a novel scheme for mining the state association rules of process object. The method includes *k*-means algorithm, association rule mining algorithm and other technologies. The rules which contain a close relationship of the links bring the significance. Firstly, it helps decrease the cost of production and enhance the productivity. The association rules show the hidden relationships, so that we can directly increase or decrease some links to adjust the output links instead of waste time on the others. Secondly, from the abnormal rules, some failure can be detected. Knowing well about abnormal states can help to find the reason quickly when the failure happens. So we can gain some knowledge and give the process industry some industry guidance to improve the efficiency of production and operating.

However, the results may be not ideal. The percentage of the transition part is a little high. At the same time, the kind of the state is more than expected. From that point of view, we will try to improve the algorithm of determining the state of each link in every moment.

References

[1] Fu, Z.H. (2014) The Development and Characteristics of Big Data. http://www.leiphone.com/news/201410/NgTsZw3yDjEbk9on.html

[2] Han, J.W. and Kamber, M. (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann Pulishers, Inc., San Francisco.

[3] Tan, P.N., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. Pearson Education, Inc., London, 30-336.

[4] Rosenstein, M.T. and Choen, P.R. (1998) Concepts from Time Series. *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, Madison, July 1998, 739-745.

[5] Song, C.Y. and Ge, T.J. (2013) Discovering and Managing Quantitative Association Rules. *Proceedings of the* 22*th ACM International Conference on Information & Knowledge Management*, San Francisco Bay Area, 2429-2434.

[6] Nanopoulos, A., Papadoulos, A.N. and Manolopoulos, Y. (2007) Mining Association Rules in Very Large Clustered Domains. *Information Systems*, **32**, 649-669. http://dx.doi.org/10.1016/j.is.2006.04.002

[7] Yin, S.S., Ma, Z.Q. and Mao, W.D. (2012) Association Rule Algorithm Based on Chain Storage Method of Frequent Itemsets. *Computer Engineering and Design*, **33**, 1002-1007.

[8] Zhu, Y.Q., Sun, Z.H. and Ji, X.J. (2003) Incremental Updating Algorithm Based on Frequent Pattern Tree for Mining Association Rules. *Chinese Journal of Computers*, Sciences Press, Beijing, 91-96.

[9] Zhou, L.J. and Wang, X. (2014) Research on Association Rules Algorithms Based on Cloud Environment. *Computer Engineering and Design*, **35**, 499-503.