Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Research on Big Data Cluster Analysis Method of IoT

Lei Chen

College of Electrical and Information Engineering, QuZhou University, QuZhou City, Zhejiang Province, China

wzuccieei@126.com

Abstract: In With the rapid development of Internet of Things technology, there have been many applications related to the Internet of Things. While these applications have brought great convenience to people's lives, they have also produced a large amount of data. The effective discovery of valuable model relationships from the Internet of Things big data helps managers to make correct decisions about the company's future development trends. At the same time it is also conducive to improving corporate profits. This article starts with the analysis of complex event relationships and processing techniques to study the analysis methods of Internet of Things big data. After introducing the concept of complex event relations, the big data processing of the Internet of Things has been transformed into the extraction and analysis of complex relationship patterns, thus providing support for simplifying the processing complexity of Internet of Things big data. The analysis of big data by clustering algorithm helps us to find the potential correlation within the data.
Keywords: Internet of Things, Complex event, Big Data, Clustering Algorithm

---

1. INTRODUCTION

IOT[1] (Internet of things) refers to the combination of various hardware devices and the Internet through information sensing devices to form a vast interconnected network of objects and things; the sensing devices include RFID devices and infrared sensors. Device, laser scanner, two-dimensional code, and so on. Among them, RFID sensor technology is widely used in various application fields. The analysis of Internet of Things big data can be analyzed based on the complex event[2] flow formed by RFID. It mainly uses the CEP technology to form complex event relationships with RFID, according to the data source. It establishes an event relationship model and then clusters the resulting event relationship model.

2. ALGORITHM DESIGN

Clustering algorithm is currently the most widely used data mining algorithm. Its main research areas have been widely used from pattern recognition to data analysis and image processing. Especially for the processing of Internet of Things big data, its excellent performance, widely loved by the academic community, and has also been widely used in the industry. Aiming at the heterogeneity and

multi-dimensionality of the data of the Internet of Things, how to select an appropriate clustering algorithm has become one of the important technical problems for researchers in this field.

EPCglobal is an RFID standards research organization. It has been 10 years since its initial release. Its main responsibility is to formulate global standards for EPC networks to form a unified standard that will facilitate faster delivery in the supply chain. Effectively identify products automatically and provide basic functionality that meets all of the company's application cases. RFID technology can be used for information exchange in EPCglobal's network, and EPC networks mainly use EPCIS specifications to implement EPC-related shared data between enterprises.

In the process of handling the RFID event flow, the definition of the event is first analyzed in detail, and then through the definition of the relationship between the events in the event flow, a specific definition of the relationship between the relevant patterns is analyzed. The relationship between events was quantified based on the relationship between events. Cluster analysis[3] was performed by quantifying the size of the distance.

A cluster analysis algorithm is now designed to perform clustering on the dataset to see if the clustering results approximate the proportional distribution generated by the simulation dataset. As the result of clustering will be different due to the different selection of clusters, different results will appear. In a specific scenario, multiple experiments can be performed, and the average result value of cluster results after multiple experiments is used as the experimental result. Reference object. Specific to the application scenario, the congruence relationship may be determined based on the time attribute and the site attribute; the synergistic relationship may be determined according to the event type and location attribute; and the causality is determined based on the commodity class attribute, the time attribute, and the site attribute through the above clustering attributes. Select and obtain the corresponding event relationship, and then quantify according to the complex event related attribute values, so as to perform cluster analysis on the data set and obtain an effective event pattern relationship. The specific relationship is as shown in Figure 1.

The main method used by the K-means algorithm is a partition-based clustering method. The core of the K-means algorithm is: for the selected data set, k points in the data set space are selected as the clustering center, and the clustering operation is performed. By the mathematical method of Euclidean distance calculation, the objects closest to them are attributed to the same class. Then, the clustering process is repeated continuously, and the value of the center point is continuously updated during the process until reaching the set number of repeated iterations or exceeding the set boundary value of the rule function, to obtain the best clustering result. However, in the above K-means clustering method, it is necessary to artificially intervene to set the value of k. Here, Dijkstra's algorithm can be used to first obtain the Canopy number and use it as the value of k in the K-means clustering algorithm. Based on the setting, this will reduce the blindness of selecting k values to some extent, and then use K-means algorithm to perform the final clustering of data.

The algorithm is described in pseudocode as follows:

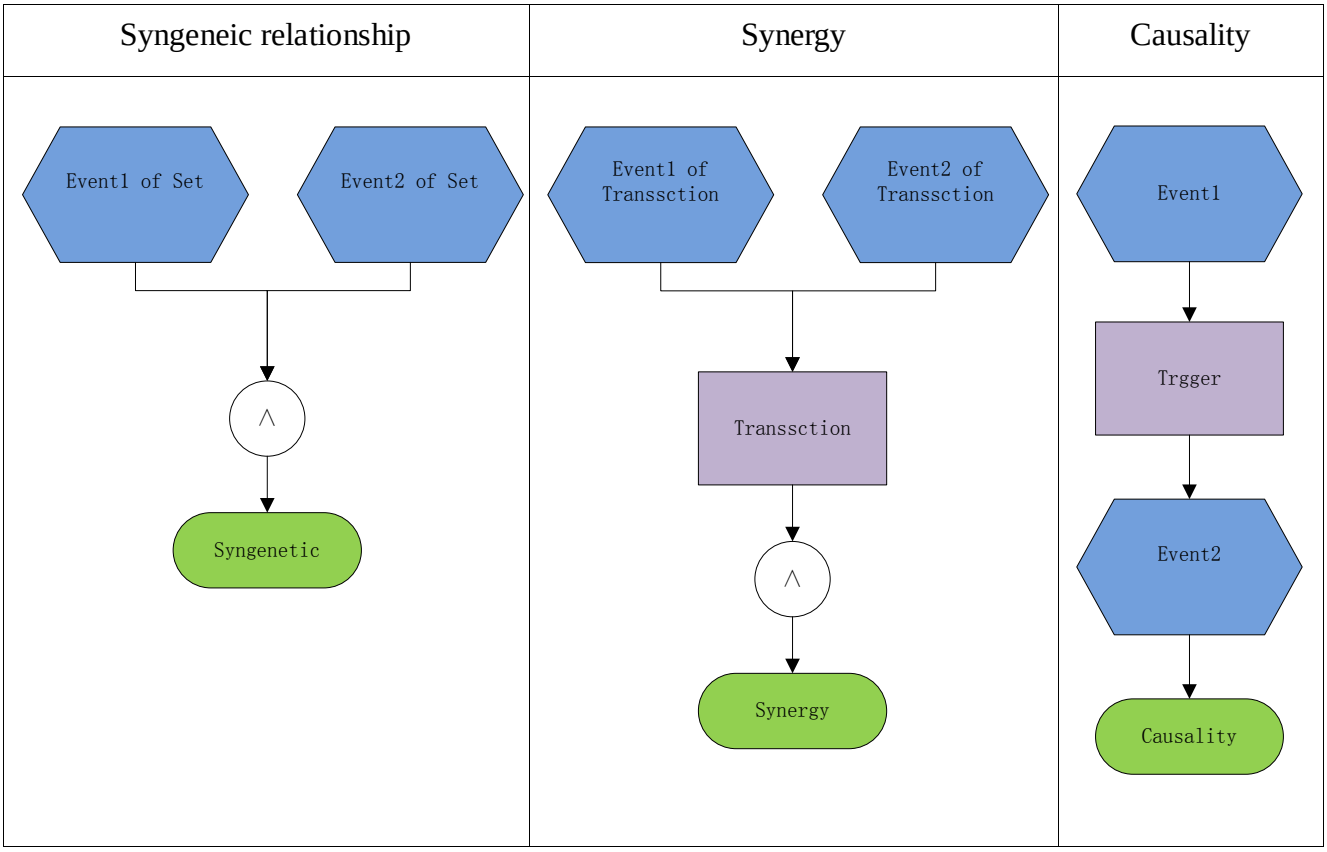

Fig. 1 Relationship between pattern relationships and event attributes

Algorithm1 Canopy-kmeans algorithm for event patterns based on CEP model

Input:Data:List(D),T1,T2

Output:The result of the clustering algorithm

Begin

Long k=0;

While(D !=Null)

{

Take the first element of D as a cluster center;

While(D did not traverse completed)

{

The remaining elements of the enumeration D,The current value is X,Calculate dis(x, center)

dis<T1,center.add(X);

dis<T2,D.remove(X);

}

centerList.add(center);

k=k+1;

}

Get input data W;

While(Algorithm does not converge)

{

For the K center point, data set centerList;

Find all data points that fall into this category;
Change your own coordinates to the coordinates of the center point of these data points;
}
end

In algorithm 1, D is a data set, T1 and T2 are two thresholds respectively, and k is the number of clusters in the data set. The final result is stored in the form of Map<key, value>; where key is an array of long integers, the cluster number of the cluster to which the data belongs is recorded, and the value is a string of type String, and the data in the corresponding dataset is recorded.

3. EXPERIMENTAL RESULTS

In this simulation test, the experimental environment shown in Table 1 was used to perform cluster analysis on data sets of different sizes. This experiment is based on an e-commerce company in the warehouse through the RFID reader to identify and record items to form a certain flow events. Based on this real data set, data sets spread at different scales are fixed according to a certain ratio in order to make reference to the accuracy and efficiency of the experimental results. In this experiment, 20 EPC read points were set. Create vertical data streams.

Table 1 Experimental environment parameter table

Environment	CPU	Operating System	Platform
Cluster	2.4GHz	Linux	Hadoop-2.6.1
PC	3.4GHz	Windows	Eclipse Luna(4.4.0)

In order to test the effectiveness and operational efficiency of this experiment, simulation tests were performed on Table 2 (simulation data set) through five groups of experiments. This data set was based on preliminary analysis of the actual data collected by the event data collection system. Based on this, the expansion is generated at a certain rate.

Table 2 The structure of the simulation data set

Data(GB)	Relationship	Number of models
0.5	Syngeneic	500000
	Synergy	200000
	Causality	300000
5	Syngeneic	5000000
	Synergy	2000000
	Causality	3000000
10	Syngeneic	10000000
	Synergy	40000000
	Causality	60000000

For the simulation data set, the canopy algorithm was used to determine the optimal clustering number of data sets. Because of the complexity and uncertainty of the cause and effect relationship, when considering the determination of cause and effect, only for the event sets belonging to the same problem sub-domain, the data set is first subjected to "similiar relationship" clustering, and then each

The data sets in clusters are clustered according to the distance between the event type attribute values, so that a "cause and effect" relationship exists between different clusters in the clustering results. In order to ensure the accuracy of the experimental results, in this paper, the same experimental data is taken for multiple calculations to obtain the average value as a reference value of the experimental data, and the time used for each execution is also recorded. Therefore, the experimental results of the time efficiency simulation data set of the algorithm can be obtained directly (average of 20 clustering results) as shown in Table 3:

Table 3 The influence of the K value obtained by the clustering algorithm on the result

Data(GB)	Relationship	Time(s)	Accuracy
0.5	Syngeneic	19.708	89.95%
	Synergy	19.225	88.81%
	Causality	20.275	93.07%
5	Syngeneic	298.221	88.05%
	Synergy	971.695	90.14%
	Causality	856.294	96.05%
10	Syngeneic	46345.86	87.21%
	Synergy	10643.66	94.86%
	Causality	284332.8	91.93%

The corresponding clustering result is based on the size of the data set. The change in the accuracy of the pattern relationship is shown in Figure 2:

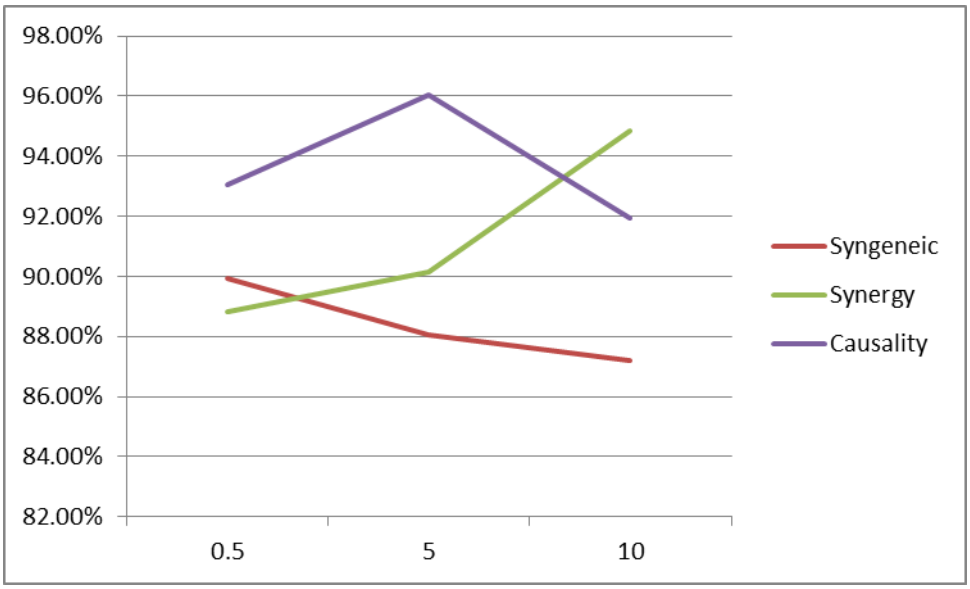

Fig. 2 Changes in accuracy of pattern relationships

In this paper, by selecting the attributes of complex events in the dataset and quantifying the attributes, we can effectively find out the relationship between the event patterns hidden in different datasets. The algorithm shows a good stability and can illustrate the feasibility and effectiveness of the algorithm. . However, due to the limited experimental environment, when the size of the data set exceeds 10 GB, the experimental environment will crash or the time required for clustering will be

outside the controllable range, and the ideal implementation result cannot be obtained. It needs to be further optimized.

4. CONCLUSION

For the characteristics of the clustering algorithm itself, when the value of k is selected, the participation of human factors will lead to the instability of the cluster analysis result. Therefore, Canopy algorithm is used to estimate the k value by the clustering center point. To a certain extent, the quality and stability of the clustering result are improved. Through the analysis of the clustering results, we can realize the needed model relations from the complex events of big data. The experimental results show that the proposed algorithm is feasible and has certain practical significance.

ACKNOWLEDGEMENTS

This paper was supported by University Laboratory Research Project of Zhejiang Province (YB201721).

REFERENCES

[1] Lee I, Lee K. The Internet of Things (IoT): Applications, investments, and challenges for enterprises[J]. Business Horizons, 2015, 58(4): 431-440.

[2] Park H, Hsiao E, Piper A. Continuous query language (CQL) debugger in complex event processing (CEP): U.S. Patent 9,329,975[P]. 2016-5-3.

[3] Assunção M D, Calheiros R N, Bianchi S, et al. Big Data computing and clouds: Trends and future directions[J]. Journal of Parallel and Distributed Computing, 2015, 79: 3-15.