Research on Distributed Data Stream Mining in Internet of Things

XU Liancheng
School of Information Science and Engineering
Shandong Normal University
Jinan, China
e-mail: xusdnu@163.com

XUN Jiao
Shandong Provincial Key Laboratory
for Distributed Computer Software Novel Technology
Jinan, China
e-mail: xunjiao1987@126.com

***Abstract*—For mining useful data from mass data generated by Internet of things, analyses shortages of the traditional Apriori algorithm which has a lower mining efficiency and occupies the larger memory space. So, MapReduce model of cloud computing is introduced. In the mechanism of MapReduce, combine the architecture characteristics and key technology of Internet of Things, conduct distributed-mining on data and information in environment of the Internet of Things, and the calculation model of distributed data stream mining is drawn. Performance analysis proves that the new data stream mining model overcomes difficulty of traditional data mining, and each link reflects the idea of distributed structure and improves mining efficiency obviously.**

Keywords—Internet of Things; cloud computing; MapReduce model; distributed-mining; data stream; data mine

I. Introduction

With the rapid development of the networking industry, more and more End-users access to the network, the huge amount of data make it more difficult to find useful knowledge. Since the data generated by Internet of Things have the features: large numbers and highly dispersed [1], so the mining problem in massive amounts of data cannot be solved by the traditional data model. But the development of cloud computing provides a new solution [2] for this problem.

The initial concept of Internet of Things was from the Auto-ID Center which is established by MIT in 1999 and they proposed a network of radio frequency identification (RFID) system: make all items connected with the Internet by radio frequency identification and other information sensing device so that achieve the goal that intelligent identification and management [3]. The real concept of "Internet of Things" was confirmed formally on the Information Society of World Summit which was held in Tunis in 2005. The ubiquitous networking communications era is coming; Internet of Things gets a new dimension of communication in the world of information and communication technologies. It makes any time and any place connected to anyone extended to connect anything, so that everything connected which is called Internet of Things [4].

There are so many researches which combined networking and cloud computing currently and most of them are always doing research around the massive data processing. But researches which combine the Internet of Things, cloud computing and data stream mining to reflect the combination of distributed mining are rarely. We study the characteristics of the data under the environment of Internet of Things deeply in this paper. We propose a new data stream mining model which is combined the key technologies of cloud computing on the basis of the existing data mining models. This model overcomes the shortcomings of traditional data mining and reflects the characteristics the massive data stream mining distributed processing approach.

II. Internet of Things and data stream

A. Brief introduction of Internet of Things

Internet of Things is a network which connects between the goods and goods and realizes intelligent identification and management of goods.

Internet of Things can be seen as a integration of information space and physical space in the broad sense, all things digital, networked, to achieve efficient information interact between goods and people between goods between man and reality environment. It makes a variety of information technology into social behavior by the new service model and achieves a higher realm of information technology in the integrated application of human society

Internationally accepted definitions of Internet of Things: through the radio frequency identification, infrared sensors, global positioning systems, laser scanners and other information sensing device, according to the agreed protocol, put up anything connected to the Internet to realize the information exchange and communication and intelligent identification, positioning, tracking, monitoring and management [5].

From the point of the structure of Internet of Things, it is divided into four layers basically: sensing layer, transport layer, data layer and data control layer [6].

(1) Sensing layer [7]: to achieve the identification of good and collection of data by the sensor mainly. That is some of the same or heterogeneous sensors perception test objectives cooperatively.

(2) Transport layer: The main achievement is to transmit

and delivery information. It transmits the information collected from the perception layer to the terminal to analyses and process. It's Internet-based network is the Internet and it can be a specific industry network. The technology which is related to network layer: Distributed data processing technology, such as data stream mining technology, cloud computing, data fusion technology etc.

(3) Application layer: Include middleware technologies, based on cloud computing the application processing center and system integration technology.

B. The calculation mode in the Internet of Things

The calculation mode in the Internet of Things is two kinds the cloud computing model and physical computing model. Only combine these two kinds of modes organically, it can achieve the required calculation, control and decision-making.

(1) Cloud computing model:

Cloud computing model is to collect the data from the Internet of Things by a distributed architecture. Then collect and process the information using the method that cloud computing model set above.

(2) Physical computing model:

Physical computing model is based on the embedded, emphasis on real-time control; it has a high performance requirement for the terminal equipment. This intelligence is embedded, is the use of intelligent information processing results. It cannot be established on the basis of the complex calculation of the terminal and the requirements for centralized processing power and system bandwidth are relatively low.

III. Data stream frequent pattern mining in the environment of Internet of Things

A. Difference between data in the environment of Internet of Things and traditional data

With the continuous development of networking technology, it is formed a form which is different from the conventional static database data - the data stream in various fields. For example, phone records data stream in the communication field, data stream of user clicks on the Web, the network monitoring packet stream, various types of sensor networks detection data streams, the Securities data streams in the financial sector, image data stream returned by satellite and retail business transaction data stream, etc. People often tend to get useful information regardless of time and place. These data and traditional data are different, the difference in Table 1. So, how process the data streams timely and efficiently and to dig out useful knowledge become a new topic of data mining in Internet of Things environments.

TABLE I. Difference between data in Internet of Things and traditional data

Comparisons	traditional data	data in Internet of Things
Dynamic	Static data	Constantly updated dynamic data
Growth rate	Very slow	Exponential growth
Dimension	One-dimensional	Multidimensional
Data processing	Offline processing	Online real-time processing

B. Frequent pattern mining based on data stream

Definition 1: data stream

The data stream is a sequence of data items reaches a certain speed in a continuous x_1, ..., x_i, ..., x_n, ... The sequence of data items only be read once by subscript i in ascending order.The data stream has characteristics that real-time, continuous, orderly, quickly on-line analysis, unpredictable.

Definition 2: Frequent item set

If the support of item set is greater than or equal to the pre-defined minimum support threshold, then the set is a frequent item set [8].

Nature 1: if the K-item set $\{I_{i1}, I_{i2}, ..., I_{ik}\}$ support number is not less than the minimum support number. Then $\{ I_{i1}, I_{i2}, ..., I_{ik} \}$ is a frequent item set.

Definition 3: Support number and support

The number of times that a set appears in the data of the transaction centrally is called support or support count number. Ratio of support number and the total number of transactions is called the set of support.

C. Apriori algorithm and its problems

Apriori algorithm [9] uses an iteration which is called layer by layer search method. K-item set is used to explore (K+1)-item set. Firstly, by scanning the database, accumulate the counts for each item, and collect items meet the minimum support, to find frequent 1-item set collection, the collection is denoted L_1.Then, L_1 is used to find the collection of frequent 2-item set L_2. L_2 is used to find L_3, …, until can not find k-item sets. L_{k-1} obtained L_k need the process of connecting two-step and the pruning step process totally.

Assume that the length of transaction database is n, to generate a maximum length of frequent item sets of m, the Apriori algorithm should scan on the database m times, its time complexity is $\Omega(n\times m)$. And the Apriori algorithm will generate a lot of candidate sets, caused a waste of spaces.

IV. Platform of Internet of Things environment - cloud computing platform

A. Introduction of cloud computing

Cloud computing was proposed by Google in 2006 firstly, then IBM [10] Amazon [11] launched its own cloud project, promote the continuous development of cloud computing.

The narrow sense of cloud computing is the delivery and usage patterns of IT infrastructure, refers to the network to obtain the necessary resources by the demand, and scalable way, the network provide resources are called cloud.

Cloud computing refers to service delivery and usage patterns broadly, refers to the network to obtain the necessary services by the demand, and scalable way. This service can be IT and Software, Internet-related, may be any other services, it has a very large scale, virtualization, reliable security and other unique effects.

Cloud computing is the development of parallel computing distributed computing and grid computing, or it is the commercial realization of these concepts in computer science. From the basic theory and cloud computing cloud computing model evolution point of view, the key technology of cloud computing, including massive distributed storage technology, distributed processing technology, virtualization, collaboration technologies, parallel computing techniques [12].

B. Hadoop calculation module -Map Reduce model

Parallel computing technology is the core technology of cloud computing, and one of the most challenging technologies. The core is storage module HDFS and Map-Reduce computing [13], Hadoop platform has its own distributed file system and its implementation using the MapReduce model.

MapReduce[14-15] is the core of the cloud computing model, is a kind of distributed computing technology, is also a simplified distributed programming model, which is to solve the problem of application development model, and developers also often use the mode to dismantle the problem. The architecture of MapReduce is not identical with ordinary cluster, on the high performance computer, it is the architecture [16] using the centralized storage subsystem [17] for mass data processing system.

The reason of using the cloud computing model in the Internet of things, the reason is that the cloud computing is the development of parallel computing, distributed computing and grid computing. In the Internet of things ,this kind distributed parallels is urgently needed, the Internet of things adopt the cloud computing model rather than the other distributed parallel computing model.

C. The relationship between the Internet of things and cloud computing

Under the environment of Internet of things ,the advantage of cloud computing:

(1) Low cost of distributed parallel computing environment, cheap storage space can store huge amounts of data;
(2) The development of cloud computing model is convenient, Blocking out the ground floor;
(3) The size of the data processing greatly improved;
(4) There are differences between the Internet of the demand for computing power, the good scalability of cloud computing, which can meet the demand of the difference caused by different;
(5) Fault-tolerant computing power of Cloud computing model is strong, the robustness is stronger. In the Internet of things, in the process of data collection, because physical distribution of the sensors is more extensive, so this kind of fault-tolerant computing is very necessary.

In a word, the Internet of Things links the real world objects through the sensor and the Internet, and implements cloud services by cloud storage and cloud computing. The Internet of Things depends on cloud computing, to collect, manage and storage the huge amounts of different format data, and achieve forecast and make decision.

V. DISTRIBUTED DATA BASED ON CLOUD PLATFORM UNDER THE ENVIRONMENT OF THE INTERNET OF THINGS

A. Basic framework of data stream mining based on cloud platform under the environment of the Internet of things

On the whole it can be divided into two parts : outside the cloud and inside the cloud (see figure 1).

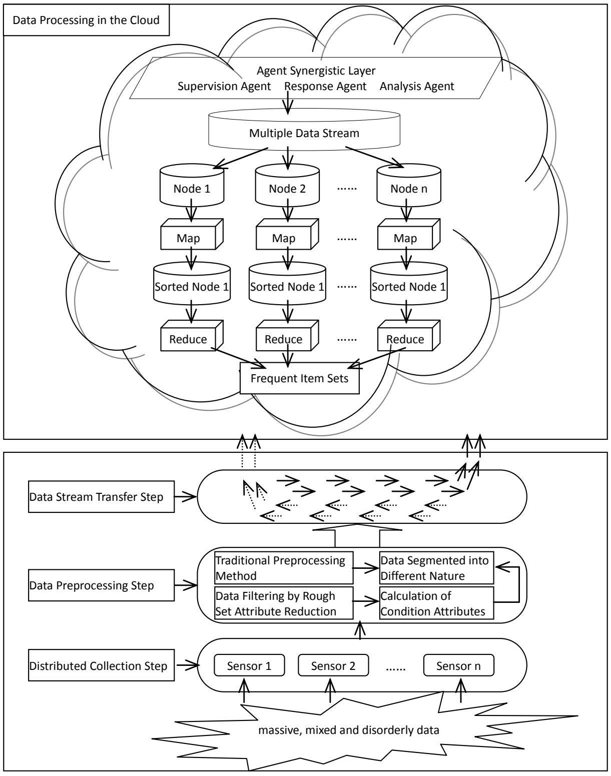

Figure 1. Basic framework of data stream mining based on cloud platform

The mining form of massive data in the Internet of Things mainly includes data stream analysis, classification and clustering analysis based on path, frequent pattern and sequence pattern analysis and outlier analysis, etc.

With the development of the Internet of things, computer and communication areas have seen continued development and data is constantly updated, as a result, a new form of data-data steam arose. Henzinger[18] and others introduced data stream as a model of data processing for the first time in 1998[19]. Therefore, the founding of frequent items in data stream under the environment of Internet of things is one of the most fundamental problems in data stream mining. Due to the characteristics of data stream, such as real-time, continuousness, ordering and quick arrival, and the application requirements for on-line analysis, many challenges have been presented to data mining algorithm, making it hard for the traditional algorithm to be applied, thus, the mining of frequent patterns in data stream based on cloud platform seems more meaningful.

Data mining is an important technical means of decision support and process control, and it is also an important part of the Internet of Things. Considering the integration of the Internet of Things and cloud platform , traditional data mining algorithms cannot solve the problem of different time in different place, while the distributed data stream mining model can do.

B. Performing Steps of distributed data stream mining model framework in the Internet of things

Combining the inherent characteristics of data stream and frequent pattern mining, on the basis of MapReduce programming model, this paper proposes a new algorithm for mining frequent pattern in data stream.

The algorithm for data stream mining is divided into five major steps on the whole.

Step 1, data collection, in which information and data will be collected in use of the perception layer of the Internet of things at any time.

Data was collected from intelligent objects of the Internet of things. The special requirements of intelligent objects need to be considered when collecting data from intelligent objects. For example, if you want to collect data from distributed sensor network, the efficiency, scalability and fault tolerance of network should be taken into consideration. A series of strategies such as regional data set can be adopted.

Step 2, data preprocessing, which is divided into two sub-steps.

(1) First of all, attribute reduction of data be conducted in use of the characteristics of rough set to filter out some data roughly;
(2) Then, reasonable segmentation of data be conducted according to the condition attribute of data.

This stage mainly reflects the concept that huge amounts of information be processed through rough set. For the segmentation of high dimensional massive data, if each attribute be scanned before the division, it will take a lot of time. Therefore, a preliminary calculation to data set must be conducted before the segmentation so as to get a set reduction; the importance of attributes of the data after reduction be evaluated and ordered; finally, the data be reasonably segmented in the use of the reduction algorithm for rough set attribute value. In this way, the efficiency of data preprocessing will be greatly improved.

Step 3, data stream transfer.

This stage includes the sending, transmission and receiving of information, finishing the task of transmitting the state of things and the way they change from a point on the space (or time) to another. The data after data preprocessing flow into the information processing layer of the Internet of things in a certain direction; and the Internet of things will make specified objects accessible to information network; relying on a variety of communication network, data or information can be transferred to the cloud platform anytime and anywhere to carry on reliable information interaction and sharing and guarantee the reliability of data from a certain sense.

Step 4, data analysis.

This stage mainly utilizing the synergistic agent mechanism of the cloud computing, service cluster integration mechanism, and multiple data stream transformation mechanism.

Synergistic agent mechanism of the cloud computing: information be distributed to the cloud server through the transport layer of the Internet of things ; after that, the cloud server will make data nodes which spread physically constitute a distributed system logically. Active coordination agent with the characteristics of active migration be used for the migration and coordinated mining between distributed nodes .Considering so many nodes in the Internet of things , calculation can be distributed on each node; Agent will actively release back a small amount of result information after the mining and induction of each node to avoid the storage and transferring of large data and scattered computation, which fully shows the superiority of distributed data mining technology.

Service cluster integration mechanism: under the Internet of things based on cloud platform, real-time management and control of massive information within the network be conducted through a group of central computers with supercomputing ability, to provide a good user interface for the upper application.

Multiple data stream transformation mechanism: as the Apriori algorithm is only for single data flow, it is a necessary step to convert multiple data flow into single data stream, and this would require the use of MapReduce mechanism.

Step 5, decision-making processing and rule generation.

This stage mainly applies the MapReduce model mechanism of cloud computing, the reduction mechanism of rough set theory, and the improved data mining algorithm.

In the process of data mining, the processing of decision-making and the generation of rules can be obtained through association rules mining which is an important algorithm for data mining ; under environment of the Internet of things, the Apriori algorithm be firstly improved in terms of performance and there are many methods for its improvement. Then, the improved Apriori algorithm should be Map/Reduce modeled in use of MapReduce model in the cloud; frequent item sets and “pure” frequent item sets be generated or obtained on the basis of applying the rough set attribute reduction algorithm to further reduce data and form reduced data set, which can fundamentally solve the problem of extracting all the useful information from large amounts of data.

VI. Performance analysis

Compared with the traditional data mining algorithm, the efficiency of the new data mining algorithm model proposed in this paper is shown as follow.

(1) In the step of data collection: the new algorithm makes full use of the working mechanism of perception layer of the Internet of things , directly collects data from the distributed sensor network anytime and anywhere, and is more intelligent than the traditional data mining algorithm.

(2) In the step of data preprocessing: it has two more steps than the traditional data mining algorithm, the first of which is data reduction using rough set theory, the second is data segmentation. After these two steps, useless information in huge amounts of data can be got rid of and the massive data in the Internet of things can be processed more effectively.

(3) In the stage of data transmission: the new data mining model for data transmission is performed in the form of data stream, which further reflects the orientation and dynamic of data. This new algorithm is more suitable for dynamic data mining.

(4) In the stage of data analysis: for the Internet of things has the characteristics of possessing so many nodes, synergistic agent mechanism of the cloud computing and service cluster mechanism of super computer with high performance are applied to avoid the storage and transferring of large data and scattered data calculation, which fully shows the superiority of distributed data mining technology.

(5) In the stage of decision-making processing and rules generation: traditional data mining algorithm only operates frequent item sets mining algorithm in the general environment; while the new algorithm model applies MapReduce model mechanism in a cloud environment, combines the reduction mechanism of rough set theory and improved data mining algorithm,gets streamlined frequent item sets library ,through which useful rules are generated.

VII. Conclusion

Cloud computing is the cornerstone of the Internet of things, and data mining is a necessary part of the Internet of things. If the intelligent information processing and data mining are not added, the Internet of things is just a sensor network and can't reflect the characteristics of intelligence. Based on the massive data storage platform of cloud computing and the MapReduce model, this paper provides a new mining platform , new methods and a new processing method for the mining of huge amounts of data stream under environment of the Internet of things, completely solves the problems of distributed storage and distributed mining of massive data in the traditional data mining algorithm. However, in terms of network efficiency, extensibility and fault tolerance of the Internet of things, the implement of the new algorithm will increase more cost than the traditional data mining. Under the big environment of the Internet of things, how to protect and process the complicated data with low cost, high intelligence and enough privacy will be studied in the future.

Acknowledgment

Thanks to the anonymous reviewers. This work is supported partly by National Nature Science Foundation of China (60873247), Science and Technology Plan in Colleges and Universities of Shandong Province (J12LN21), Nature Science Foundation of Shandong Province (ZR2011FM030).

References

[1] Yu Fanpeng, Niu Yanchao. Evaluation of Hop-distance Relationship for Sensor Networks in Internet of Things [J]. Computer Science, 2012, 39(3): 107-109(in Chinese).

[2] Wang Peng. Cloud Computing [M]. Beijing: Posts and Telecom Press, 2009(in Chinese).

[3] Sun Qibo, Liu Jie, Li Shan, et al. Internet of Things: Summarize on Concepts, Architecture and Key Technology Problem[M]. Journal of Beijing University of Posts and Telecommunications, 2010, 33(3): 1-9(in Chinese).

[4] International Telecommunication Union Internet Reports 2005: The Internet of things[R]. Geneva: ITU, 2005.

[5] Hu Xiangdong. Survey on Research and Development of Internet of Things[J]. Digital Communication, 2010, (2): 17-21(in Chinese).

[6] Wu Zhenqiang, Gao Yanwei, Ma Jianfeng. A Security transmission Model for Internet of Things[J]. Chinese Journal of Computers, 2011, 34(8): 1352-1361(in Chinese).

[7] Joydeep Ghosh. A Probabilistic Framework for Mining Distributed Sensory Data under Data Sharing Constraints [C]. First International Workshop on Knowledge Discovery from Sensor Data, 2007.

[8] Xia ying, Zhang jun, Wang Guoyin. Spatio-temporal Association Rule Mining Algorithm and its Application in Intelligent Transportation System. Computer Science, 2011, 38(9): 173-176(in Chinese).

[9] Chen Anlong, Tang Changjie, Tao Hongcai, et al. An Improved Algorithm Based on Maximum Clique and FP-Tree for Mining Association Rules[J]. Journal of Software, 2004, 15(8): 1198-1207(in Chinese).

[10] SIMS K, IBM introduces ready-to-use cloud computing collaboration services get clients started with cloud computing [DB/OL], http://www-03.ibm.com/press/us/en/ressrelease/22613.wss, 2007.

[11] Amazon, Amazon elastic compute cloud (Amazon EC2) [DB/OL], http://aws.amazon.com/ec2/, 2009.

[12] Chen Kang, Zheng Weimin. Cloud Computing: System Instances and Current Research[J]. Journal of Software, 2009, 20(5): 1337-1348(in Chinese).

[13] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters//Proc of the 6th OSDL. SanFrancisco: USENIX, 2004: 137-150.

[14] Dan Weld. MapReduce: Simplified data processing on large clusters. 2011-07-03]. http://rakaposhi.eas.asu.edu/se494/notes/s07-map-reduce.ppt.

[15] Ghemawat S, Gobioff H, Leung S T. The google file system// Proc of the 19th ACM Symp on Operating Systems Principals(SOSP). New York: ACM, 2003:29-43.

[16] Xie Xianghui, Peng Longgen, Wu Zhibing, et al. Research on High Performance Computer Technology Based on InfiniBand[J]. Journal of Computer Research and Development, 2005, 42(6): 1-12(in Chinese).

[17] Zhao Yi, Zhu Peng, Chi Xuebin et al. A Brief View on Requirements and Development of High Performance Computing Application[J]. Journal of Computer Research and Development, 2007, 44(10): 13-22(in Chinese).

[18] Sun Yufen, Lu Yansheng. An Overview of Stream Data Mining[J]. Computer Science, 2007, 34(1): 1-5(in Chinese).

[19] Henzinger M R, Raghavan P, Rajagopalan S. Computing on data streams. SRC Technical Note 1998-011. Digital systems research center: Palo Alto, California, 1998.