Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 6 Issue 6, June - 2019

Selective Event Analysis for Efficient Data Stream Processing

Donghyok Suh[1]
Department of Display Engineering, Dankook University, Cheonan-si, Chungnam, 31116, Korea
hanhwaco@naver.com

Moonsu Lee[2]
Samduk Electrical CO., Ltd.
Kimpo-si,10117, Korea

***Abstract*—In this study, we propose a data stream processing method for efficient operation in the inspection of the condition of the switchboard and all electrical equipment by installing sensors and networks in the switchboard and all electrical facilities. The existing data stream analysis is executed in a unit of a time interval. In this study, we propose a method to divide into the parts to be analyzed and the parts that should not be analyzed. This can significantly reduce the amount of analysis for vast amounts of data streams. To determine which part to analyze, it is necessary to introduce a marker in the part to be analyzed in the data stream. At this time, the marker was based on the generation time of heterogeneous sensors. In the same situation, it was possible to guarantee the efficiency in the analysis of the data stream by setting the point of time when the different sensors detect and report the occurrence time of the event to be coincident or overlap.**

Keywords*— *IOT, Smart Distribution Panel, Data Stream, Context Awareness

I. INTRODUCTION

The emergence of the Internet of things has enabled innovative changes in the 21st-century industrial society. In the previous ubiquitous sensor network, problems of error and distance limitation in wireless data transmission were solved by the Internet of Things, a new integrated network configuration was made possible, and artificial intelligence became possible by the appearance of Internet of things. Based on the Internet of things, artificial intelligence is applied to various fields of industry, and it is triggering the innovation of the industry.

The artificial intelligence in the Internet environment is the key asset of the sensing activity of the sensor in the terminal of each network system. The performance of the cognitive activity of artificial intelligence can be enhanced depending on how good information the sensor acquires and provides. Therefore, the performance of the sensor and the processing of the data sensed by the sensor are important.

On the other hand, the data detected and reported by the sensors included in the network system continues to be transmitted over time. Data that is continuously received and transmitted without the transmission and storage is called a 'data stream'. The biggest feature of the data stream is that it is continuously available and the sum of the data is very large. Processing new forms of persistent data has become an important issue in the recent data processing.

This study is based on a smart switchboard system. This paper proposes a method to improve the quality of the situation information while reducing the amount of computation in processing the data stream generated and transmitted in the environment where the sensor and network are installed in the distribution board and related electric facilities. It is often difficult to standardize and store the data stream. The existing method of analyzing after completing the storage is difficult to use for the current data stream. Since the data stream often does not limit the time available, the total amount of data accumulated over time continues to increase. Therefore, in data stream analysis, reducing the amount of computation is an important issue. However, it is helpful to have a wealth of data for recognition to recognize the actual situation you are aiming for without human intervention.

To reduce the amount of computation in the data stream analysis and to obtain the quality data necessary to acquire the situation information may be contradictory or in conflict with understanding. The ideal data stream is the data type that is inherently continuous and variable. Failure to reduce the amount of computation in the ideal data stream analysis results in a large cost in processing the data stream, which is an issue for ensuring the efficiency of data processing.

Therefore, it is important to identify what to process and what to do in the data stream processing. In this paper, we propose to create a marker on the data stream for the identification interval to achieve the goal of identifying the segment to apply the analysis algorithm and the segment that does not need to ensure efficiency in the data stream processing. It suggests whether it is. Obtaining an indicator on the data stream can be difficult but important. Creating an indicator for analysis in a data stream is difficult because the data stream is not a material flow, and it can be difficult to create an artificial marker because the data is not due to artificial generation but rather to sense the real world phenomenon. However, selective and rational data processing is inevitable, rather than continuing to analyze all data streams for continuously and infinitely inflowing data streams. We used the characteristics of the data stream to set reasonable indicators on the data stream to detect and report the phenomena that accompany real-world situations.

This paper is organized as follows. In Section 2, related researches are summarized and in Section 3, proposed a method to set the analysis start indicator for efficiency of data stream processing. In Section 4, experiments and evaluations are conducted and conclusions are made in Section 5.

II. RELEVANT STUDIES

A. Real time data stream processing

Stream data collected through the sensor network requires real-time processing and occurs continuously and continuously. There is a problem that the stream data is large and difficult to store, and it takes a lot of time to retrieve the data.

TinyDB, a real-time sensor network query processing system, is based on the concept of database tables and limits energy efficient distributed query processing methods for efficient query processing. Using Tiny-SQL, we can efficiently increase the usage of TinyDB query language[1]. All the data obtained through the sensor nodes are filtered through the query sent from the user at each sensor node rather than coming up to the user. This reduces energy consumption by optimizing the transmitted data in a way that saves the energy consumed in transmitting the data[2]. To reduce unnecessary metadata search and processing process in a sensor network environment, only the necessary metadata is configured in cache table, and the performance of data processing system is improved with minimized access to database[3]. The heterogeneous sensor stream transferred from the sensor network is converted into context-aware data and stored in the storage. For this purpose, the situation stream processing system manages the catalog information between the sensor streams to convert the sensor data into context-aware data and to combine the information of each sensor stream and the situation information. It also defines continuous queries as boolean functions and provides various operators, continuous queries. The proposed situation stream processing system consists of a sensor stream catalog manager, sensor stream analyzer, sensor stream continuous query manager, and sensor stream converter [4]. To improve the real-time reactivity in the IoT environment, a comparative analysis of stream data processing methods and a data processing tool for fast-reaction were analyzed. K-means clustering is a method of dividing data into k clusters. The distance from all the data to the center k is obtained, and each data is included in the closest cluster. The K-means machine learning method collects data using the Collect class to generate a cluster model. We created a model that divides the collected data into clusters, classifies the data using a model created by the K-means machine learning method by receiving data in real time through Apache spark streaming. FastData analysis method for analyzing various stream data is processed in conjunction with Apache spark [5].

There are many restrictions on the application of the distributed system technique for real-time data processing. Apache Spark and Strom are being developed for hierarchical data stream processing. We need to study algorithms for real-time data processing [6].

B. Switchboard Safety Inspection System

Accidents in the switchboard are caused by insulation failure, natural deterioration, and overload due to various stresses during the short and long term. When the electrical equipment in the switchboard is faulty, PD measurement and temperature measurement are performed simultaneously. Transformer failure is detected by a local temperature rise and PD test. The series reactor capacity is designed to be smaller than the capacitor capacity, and the overcurrent flows to overheat. In the PD test, only noise with no special pattern in the signal is detected. The cracks in the support of the BUS-Bar are difficult to distinguish by the naked eye. In the PD test, a certain pattern of a signal can be detected and the failure can be detected [7].

Petrochemical complexes are more vulnerable to damage than small industrial accidents because they can lead to a series of explosions. As a method for supplying safe electric equipment in petrochemical complexes, there is a deterioration diagnosis method in water and substation facilities. We compare the temperature difference by the thermal imaging camera with the deterioration diagnosis of the water and power supply facilities. Analyze the deterioration diagnosis by photographing the switchboard and outdoor substation with a digital camera and thermal camera. It is necessary to study chaos theory and the automatic discrimination technique by signal processing [8]. It is possible to diagnose various electric safety conditions such as deterioration state, defects, and the electrical connection state of terminals by measuring the temperature distribution signal generated in the electric facility in real time. The number of electrical equipment for hospital medical equipment, the deterioration condition of electrical equipment was diagnosed after telescopic measurement of the surface temperature distribution image using the high-performance IR camera of the incoming cable, breaker, insulator, and power cable of the substation facility [9].

This is a method for monitoring the switchboard using a non-contact infrared temperature sensor. The infrared temperature sensor measures -20 to 300 ° C and the ratio measures the power device at an insulation distance of 15: 1. We calculated the conversion equation for the bus bar through the characteristic experiment and applied the DAU to compensate for the error. Monitor the temperature change of the measurement object according to the ambient temperature and observe the abnormal symptoms [10],[11].

III. SEMANTIC-ORIENTEC DATA STREAM PROCESSING FOR CONTEXT RECOGINTION IN SMART SWITCHBOARD

The data stream covered in this study is data that continuously transmits the values detected by various kinds of sensors to the host. Data analysis for situational awareness should be done. At this time, it may be difficult to analyze all the data of the stream

type. The reason why it is difficult to analyze all the data streams is as follows.

A. The amount of data to be processed is large and continues to increase

In the data stream environment, the data that flows in and is reported to the host continues as long as the sensor operates. It is hard to say that it is an efficient to use of information assets to process all incoming data continuously. There is a need to select the amount of data to be processed to perform efficient processing. If so, which data will be processed and which data will not be processed? Which data is heavily processed and which data is simply processed? These issues should be considered. Continued analysis of all incoming data will be a very difficult problem in the long run. It is possible to process continuously incoming data only when a limited number of sensors are in operation. However, assuming that the number of used sensors increases and the number of types increases, the amount of data. The processing assets must also be upgraded to meet the requirements. After all, these costs are not endless. Therefore, the selective processing of the data stream is inevitable.

B. An intensive analysis is necessary for the part that needs to recognize the situation

The existing data stream analyzing method is a method of collectively analyzing the data stream within the time interval after setting a predetermined time interval and storing the result. In the event of an unusual situation or an urgent situation, a very fast and accurate response is needed when there is a sign or indication of such a special situation. Rather than applying a batch time interval for all data streams and performing a batch analysis at each time zone, it is necessary for a particular portion of the data stream to follow the weight of the analysis task. This is because the smart switchboard does not only monitor the stability of electricity supply and demand for the customer but should also infer the risk of fire, explosion, etc. caused by abnormal electrical safety and electrical equipment. Accidents such as fire, explosion, and electric shock should be avoided or recovered once it is difficult to recover, so it is important to take preventive measures and to take prompt and serious attention to the signs related to these serious cases.

Efficient data stream processing in smart switchboards

The switchboard is where the public electrical network and the electrical circuit of the customer come into contact. Smart switchboards are equipped with a variety of sensors to monitor the status of the power supply and the consumption of electrical energy in order to maintain the stability of the switchboard itself, to analyze the data, Lt; / RTI & gt; In this environment, the data reported by various types of sensors continues to be normal, indicating that there are no major problems. It tells you that most of the contents of the data stream do not fall out of the normal range, and you do not have to analyze the data that keeps sending this state. Then you need to know where to start analyzing how to recognize where to look for attention in some normal data streams.

Artificial data generation such as communication signal processing for synthesizing and transmitting a carrier wave is difficult to consider in determining the indicator on the data stream. The sensor data is the value sent by sensing the physicochemical phenomena related to the actual world situation and the situation arising from the situation. Sensors detect and report real-world conditions, making it difficult to add artificial data. Therefore, it is advantageous to determine which part of the data stream to continue to receive, to begin the analysis, by setting the indicator using the nature and characteristics of the data stream.

Marker setting method: The role of the marker in this study is to inform that the data stream is getting along with the time and that the section where the analysis algorithm should be applied starts. If so, it is the role of the marker to pay particular attention to when the sensor results are being transmitted and to know where to start the analysis of the detection results. For this purpose, we propose to set the point where the events of different kinds of heterogeneous data streams match the markers. The indicator setting method on the data stream is schematically illustrated by the following figures.

In the conventional method, the time interval is set for the data stream continuously sensed by the sensors, and the characteristics and the attributes of the data having the limited size within each time zone are analyzed. The following figure shows that the data analysis is performed as a data analysis indicator when the occurrence of events reported by each sensor coincides with each other among the streams.

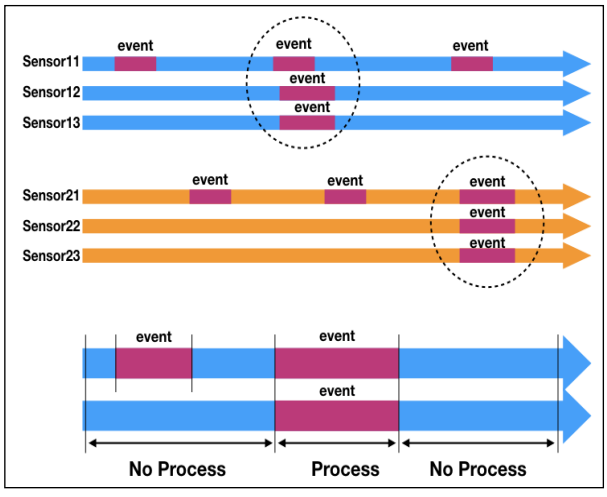

Fig. 1. Set the marker for efficient data stream processing

In the Smart Switchboard study, which is the background of this study, various sensors are installed to monitor the power state, the state of the switchboard, and the electric circuit side using the electricity. Each of these sensors performs a sensing operation and transmits a detection result. At this time, the point in

time when the various sensors simultaneously generate an event is a meaningful point. The fact that multiple sensor events are detected and reported together means that a notable change is happening. Therefore, this point should be paid attention to in the situation recognition of the smart device. When a selective and selective processing interval is determined for the data stream processing, it is reasonable to use the point where the events from the multiple sensors coincide as indicators.

The following describes an algorithm for setting such indicators.

1. $S_{n(n=1,2,3...,n-1,n)}$: Sensors

DS_1 : Data Stream of Sensor S_1

DS_2 : Data Stream of Sensor S_2

2. $Ev1_{n(n=1,2,3...,n-1,n)}$: Event data in the DS_1 $Ev2_{n(n=1,2,3...,n-1,n)}$: Event data in the DS_2

CA : Data Analysis Function for Context Awareness

3. *TC* : Time Check when The Event data appear on the Data Stream

$TEv1_n$: The Time of Event *Ev1n* occurs.

$TEv2_n$: The Time of Event *Ev2n* occurs.

4. When $TEv1_n$, *TC :* $(TEv1_n = TEv2_n)$, do *CA, TC :* $(TEv1_n \neq TEv2_n)$, do not *CA* and write the time of $TEv1_n$, $TEv2_n$.

5. Do Step 3,4 Whenever the Event occurs

Since this algorithm targets the data stream, there is no termination.

In the next section, we verify and evaluate the proposed algorithm by experiments.

IV. The Experiment and Evaluation

The following sensors were selected to verify the algorithm proposed in this study. These sensors carry out continuous sensing activity and transmit them to the host through the communication network. Here, the communication network uses the TCP / IP protocol as WiFi.

TABLE. 1 Incoming Data set (No: Number, T: Temperature(℃), H:Humidity(%))

No	1	2	3	4	5	6	7	8	9	10
T	24	23	23	23	23	24	23	23	24	23
H	29	29	29	26	29	29	26	26	26	29
No	11	12	13	14	15	16	17	18	19	20
T	23	23	23	23	23	23	23	23	23	23
H	26	29	29	29	26	26	26	26	26	29
No	21	22	23	24	25	26	27	28	29	30
T	23	23	23	23	23	23	23	23	23	23
H	26	26	26	29	29	26	26	29	29	26
No	31	32	33	34	35	36	37	38	39	40
T	23	23	23	23	23	23	23	23	23	23
H	26	26	29	26	26	26	26	26	26	26
No	41	42	43	44	45	46	47	48	49	50
T	23	24	23	23	23	24	23	23	24	24
H	26	26	26	26	26	26	26	26	26	26
No	51	52	53	54	55	56	57	58	59	60
T	23	23	23	24	24	24	24	23	24	24
H	26	26	26	26	26	26	26	29	29	26
No	61	62	63	64	65	66	67	68	69	70
T	23	23	24	23	24	23	23	23	23	23
H	26	26	29	29	26	26	26	26	26	26
No	71	72	73	74	75	76	77	78	79	80
T	23	23	23	24	24	23	24	23	23	23
H	26	26	26	26	26	26	26	26	26	29
No	81	82	83	84	85	86	87	88	89	90
T	24	23	23	24	24	11	23	24	23	23
H	26	29	26	26	26	29	26	26	26	26
No	91	92	93	94	95	96	97	98	99	100
T	23	24	24	23	24	23	23	23	23	23
H	29	26	26	26	26	26	29	29	29	29
No	101	102	103	104	105	106	107	108	109	
T	23	23	24	23	23	23	24	23	23	
H	26	26	26	26	26	26	26	26	26	

Thus, the amount of data that is detected and transmitted by smart systems is very large. The data shown in the above table shows only a portion of the incoming data stream. The data stream continues to flow and continues to report the detected value as long as the system is present. Since the data stream continues to flow, it should be analyzed differently from past data analysis methods. The time interval at which this data value was reported was 0.5 seconds and reported twice per second. The following figure is a graphical representation of the above data flow.

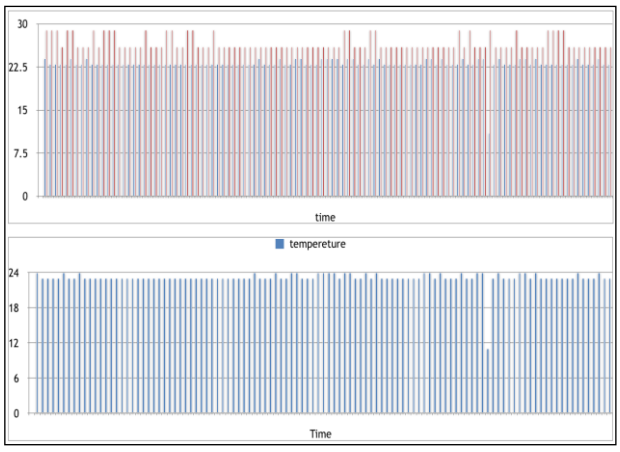

Fig. 2. Graph of incoming data

The first graph in the graph above shows the incoming temperature and humidity measurements.

The abscissa represents the time, and the time interval is 0.5 seconds, and the data flows twice per second.

In the experiment, the number of reports of various sensor devices can be adjusted. In some cases, it is set to report 10 times per second according to the experimental device. In this experiment, two measurements were sent per second. Data processing in a smart device connected to a wired or wireless network is a necessity to perform analysis on a data set continuously flowing over time. The following figure shows events that occur in the data stream.

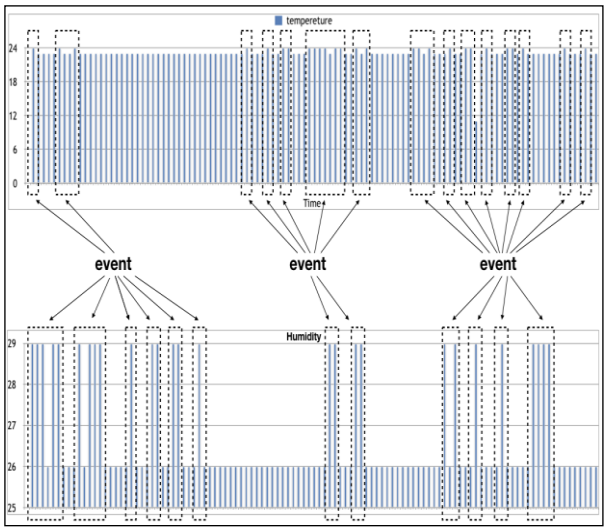

Fig. 3. Events that occur in the data stream

The upper dataset of the graph is sent by the temperature sensor. The data set at the bottom of the graph is sent by the humidity sensor. Dotted boxes are events that appear in each dataset. The following figure shows events that occur at the point of occurrence in two data streams.

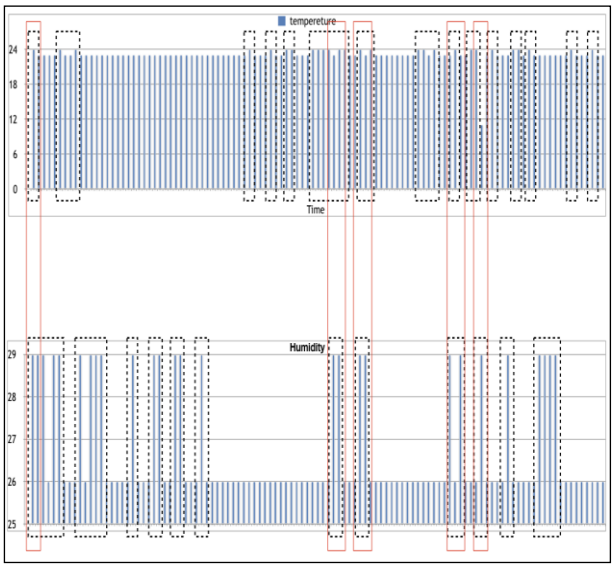

Fig. 4. Events that match the time of occurrence

The above figure shows the time when the events of the data streams detected by different sensors are matched. This study suggests that only this part should be analyzed.

Data streams are a vast amount of data sets, and theoretically, the capacity grows infinitely. Data processing in a data stream environment is difficult to target all the data that is available. Therefore, selective data processing is feasible. It is a matter of how to select the point where the problem to be solved is treated selectively. It is also difficult to create an artificial display for the incoming data. It is practically possible to find a specific property or phenomenon of the data stream and use it as an indicator. As suggested in this study, it is reasonable in terms of context awareness to use the same time as an indicator in the data stream of different sensors. If different sensors that detect and act on one phenomenon simultaneously detect and report events at the same time, it is clear proof that they have detected real phenomena.

As proposed in this study, when analyzing the data using the point at which events occur simultaneously in the heterogeneous sensor data stream as indicators, only 14.7% of the total data could be analyzed.

V. CONCLUSTION

The sensing data reported by the sensor is transmitted as stream-type data. It is inefficient to process all reported data because the data stream is continuously available. In this study, we propose a method to process the data selectively and set indicators on the data stream to improve it. Instead of generating a clean data stream that serves as an indicator, we set the same point in time at which multiple data streams in multiple sensors generate events. This is a result of common sensors detected by one sensor for a notable situation, so the host should pay attention to this and perform analysis carefully. We show that it is possible to selectively and efficiently process data streams without interruption due to the roles and functions of the indicators proposed in this study. Ensuring efficiency in data collection environments where capacity continues to increase is an important issue, so it is worthwhile to act as an indicator for selective processing in data stream processing.

REFERENCES

[1] Kim, M.G., Kim, D.H., and Kim,T.H., “SenDB : Qeury Processing System for Wireless Sensor Network”, Proceedings of the KISS Conference, Vol. 33, No. 2, pp.335–339, 2006

[2] Jang,Y.H., Lee,S.H., Kim,Y.S., and Oh,R.D.,” Query Processing System for Incomplete Sensor Stream Data of in Real-time Sensor Network”, Proceedings of the Korean Society of Computer Information Conference, Vol.22, No.1, pp.123-124, 2014

[3] Park,E.J., Byeon,J.W., Choi,D.S., Kim,J.H., and Oh,R.D.,"An Effective Stream Data Management System for the Incomplete Stream Data on Sensor Network", Proceedings of the Korean Society of Computer Information Conference, Vol.22, No.1, pp.125-126 , 2014

[4] Jovanović, Ž, "Data stream management system for moving sensor object data", Serbian Journal of Electrical Engineering, Vol.12. No.1, pp.117-127, 2015

[5] Lim,H.H., Kim,D.H., Lee, B.J., Kim,K.T., and Youn, H.Y., "Real-time stream data processing method based on IoT node cluster", Proceedings of the Korean Society of Computer Information Conference, Vol.27, No.1, pp.1-4, 2019

[6] 5Apache Hadoop,: http://hadoop.apache.org/

[7] Ko,M.S., Choi, H.K., Park, J.Y., Park,I.C., and An,J.U., "A Study on effective diagnostic equipment of switchgear", Proceedings of KIIEE Annual Conference, pp.25-26, 2013

[8] Bae,Y.C., "Degradation Diagnosis Technique of Power supply for a Petrochemical Complex using Infrared Camera", Korean Institute of Intelligent Systems, Vol.22, No.2, pp.249-251, 2012

[9] Kim,S.G.,"Thermal Image Diagnosis of Power Supply for University Hospitals using infrared Thermography", Proceedings of KIIEE Annual Conference, pp.129-129, 2015

[10] Jin,C.H., Choi, J.U., Park, S.W., and Kim,Y.G., "Monitoring and Diagnosis technique in Metal-clad switchgear by using the infrared temperature sensor", The Korean Institute of Electrical Engineers, pp.1663-1664, 2016

[11] Wang G., and Li Q., "Development of a Digital Algorithm Based on Instantaneous Power Transform for On-line Monitoring of the Dielectric Loss Factor", 2006 IEEE Conference on Electrical Insulation and Dielectric Phenomena ,Vol. 27, pp.35-42, 2007