I J C T A, 10(9), 2017, pp. 851-859 ISSN: 0974-5572

Reliable Techniques for Handling IoT Data (Unstructured Data)

V. Diviya Prabha* and R. Rathipriya**

ABSTRACT

Internet of things (IoT) is a rapidly growing area in today's world where to accomplish certain objective devices communicate over the public or private network. It acts as an interface between devices to collect or generate data without human intervention. It is also important to know the role players in IoT and their future research affairs. The first thing after sensing devices is to collect the data and integrate them to analyze data. Unstructured data is data that requires human to read data and understand for example handwritten information, audio/video dictations, email messages, machine written messages, etc., these data is converted to structured data to make a machine understand. Clustering is the one of the suitable and efficient data mining method for identifying the usage patterns from huge data collected from sensing. Being high dimensional, data mining algorithms for finding various patterns from huge web usage data has immense application like personalization, marketing, etc.,

Keywords: Internet of Things, Unstructured data, Structured data, Clustering, Data Preprocessing

1. INTRODUCTION

Data collection is an important task in business values. The gathered information or data is the effective foundation of IoT technologies. After this process managing [1] the data is the great challenge in today's environment that is storing, retrieving data is a great task. There are two kinds of data they are structured and unstructured data. From the research point of view, survey results that nearly 90% of data is unstructured data.

Structured data is stored in the relational database consist of tables or data objects. Unstructured data which is from sensors, social media, weblogs, etc., consist of audio/video, images, etc., collections of this

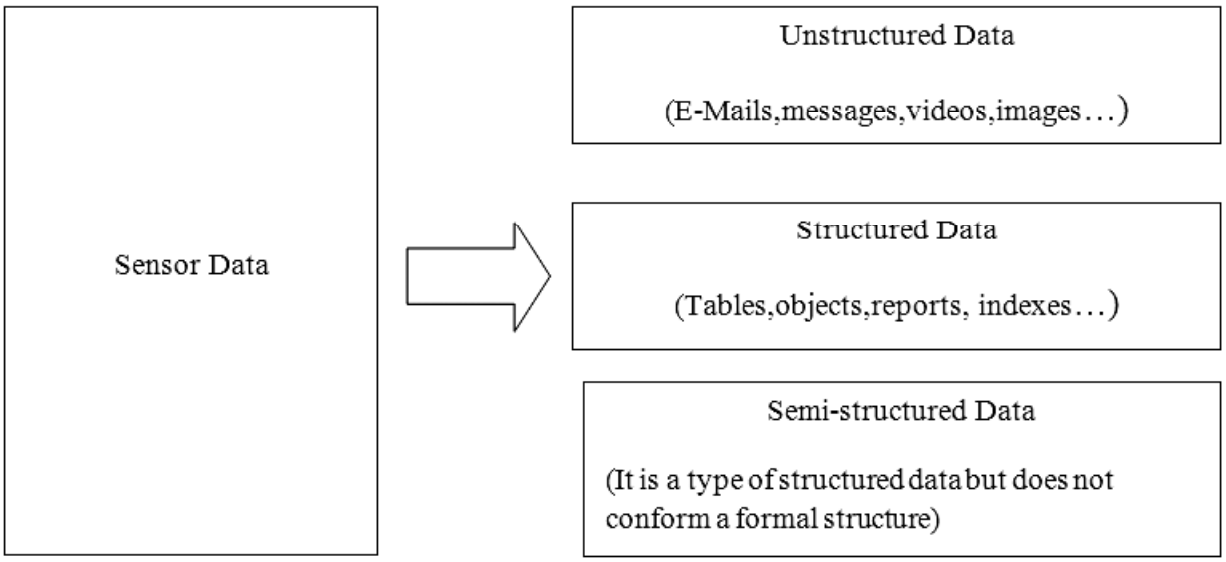

Figure 1: Structured Data and Unstructured Data

* Research Scholar, Department of Computer Science Periyar University, Salem-11, Email: diviyaprabha7@gmail.com

** Assistant Professor, Department of Computer Science Periyar University, Salem-11, Email: rathipriyarr@yahoo.co.in

kind of data is increasing day-to-day, but storage plays a vital role. In today business techniques consist of more unstructured data than structured data. The above figure represents the difference between structured and unstructured data. Most of the data collected from the IoT is unstructured data. Converting the unstructured data to structured data is an important task. Most of the data collected such as satellites images, weblogs, social media, messages, etc.

Data mining is an approach to exact useful patterns from large data. Business Intelligence technology need of data mining for business performance management, Intelligent and predictive analyses for their growth of business.Clustering techniques in data mining approaches which are used to identify similar groups or patterns. Here the data can be grouped as structured and unstructured data. In this paper, we have discussed different sources of sensors data and data management techniques. Then it overviews how to handle these unstructured data. After that, the main contribution represents the flow of converting unstructured data into structured data. Finally, clustering algorithms are discussed to identify interesting patterns three clustering techniques suggested for document data.

2. SOURCES OF UNSTRUCTURED DATA

There are different ways in unstructured collection data from sensing devices.There are two things that are to be considered in IoT

1) Data Management
2) Security challenges

In Data Management a collection of data from different sensors is managed in an efficient manner.Traditional database management system can be used to store these kinds of data. It is to collect, store and delivers the data. Unstructured data is collected and converted to structure one stored in the database system. The second task is security it is another big task for business tactics. There may be data leakage while converting and understanding of data while integrating. The following are some of the applications. The following are ways to collect the data from different sensors

2.1. Biosensing

The word biosensor was first introduced by Clark and Lyons in the year of 1962 for the term enzyme-electrode. It consists of two components bioreceptor and transducer [1] where bioreceptor recognition of target analyte and the transducer converts the recognition event into a measurable signal. So here we receive data in the form of unstructured the techniques is needed to convert intoa structured data. As the devices get smaller, it must be helpful to people in every day's life. It is a device that is used to measure biometric data. Some of the devices are wearable watches, clothing, tattoos; location tracking, etc. The following key points are helpful for biosensing

- Sensing and monitoring capacity as your wearable devices is fixed they have to sense the data and monitor it may be people, equipment, etc.,
- Management of information by the global connectivity and manage the entire devices
- The integration that is unstructured data must be transferred and integrated to structure on.

2.1.1. Blood glucose biosensor

There are many enhancing capabilities in improving the reliability of glucose measuring devices which is very much important for the blood in patients. Glucose biosensors [3] have evolved to be more reliable, rapid, and accurate and are also more compact and easy to use. Research for advanced technologies, including electrodes, membrane, immobilization strategies, and nanomaterial, continue to be performed.

2.1.2. Pulse Oximetry

It has been used for many years for patient monitoring [4] it plays an important role between patient and doctor without the doctor physical presence.

2.2. Chemical

It is used to measure chemicals in a system.The O2 sensor senses only oxygen level in the environment. Example - the Smoke detector that senses smoke and indicates fire alarm. The below is the data of smoke detector [10].

Base Module Number	:	B110LP2
Loop Type	:	2 wire
Current Limit Resistor	:	NO
Nominal Voltage	:	12/24VDC
Current Draw on Alarm (mA)	:	10-130

2.3. Current / power

Regarding the power management is one of the important sources. Based on the energy collected from outside environment the power generating element is selected. Energy harvesting is [9] important in wireless sensor based on the environment, distance and power consumption. A circuit is connected to the sensor to control the flow of current depending on the voltage level whether it is high or low voltage. Based on the voltage it is on and off state. Her data received is maintained in a structured format

2.4. Light

The light sensor is used to detect the visible light. The data has collected from the software [11] Logger Pro 3 compute program in LabQuest, the Logger Pro 2 computes the data in the serial box, etc.; Ambient Light Sensor detects the light same as the human eye, and it also can convert the light into current/voltage.

There are two ways to manage the light sensors [12] data that come from different ambient light sensors:

- Use a transform of the data so that it will be direct proportion to human behaviors. The observed give an optimal result for the solution.
- Use threshold value for a large amount of data so that it can be fixed into certain categories. These data are suitable only for the large approach.

2.5. Humidity

It detected the amount of water vapor in air or mass and measured in relative and absolute. It is based on water-phase protonic ceramic materials are used in the research lab. At that temperature what the air can hold is compared with moisture in the air. Hygrometer and humor soil moisture sensor are the examples of the humidity sensor. It has many advantages due to its cost effective and design flexibility.

Relative Humidity considers the following example [8].Data is captured from the device which does not have any format and transferred into structured format as given below:

Resolution	:	16Bit
Repeatability	:	±1% RH
Accuracy	:	At 25°C± 5% RH

Interchangeability	:	fully interchangeable
Response time	:	1 / e (63%) of 25℃ 6s
1m / s air 6s		
Hysteresis	:	<± 0.3% RH
Long term stability	:	<± 0.5% RH / yr in

2.6. Gas

Gas sensing is an important application, and its usage is increasing in industries and environment [7].It is used in different applications for example automotive industry it is used for detection of polluting gasses from vehicles.

2.7. Occupancy

It is used to detect light or heat in the space like Air Conditioning (AC) that can be on/off using infrared. It consists of passive infrared to [14] change in the temperature when someone enters the room .it can also be used in GATEWAY parking structures projects sensors must detect that light must be 24 hours per day.Is measured in low or high the heat.It can be of semi-structured form does not have a specified structure.

2.8. Position / motion

This sensor used to identify three –dimensional motion mainly used for automobiles. The above example shows the sensor of Honeywell [15]:

Series	:	103SR (digital)
Description	:	Hall-effect digital position sensor
Magnetic actuation type	:	unipolar, bipolar, latching
Supply voltage range	:	4.5 Vdc to 24 Vdc
Supply current	:	4 mA to 10 mA (inclusive)
Output type	:	digital sinking
Operating temperature range	:	-40 ℃ to 100 ℃ [-40 °F to 212 °F]

2.9. Pressure

The pressure sensor is used to measure the atmosphere pressure. For example Gauge pressure [16] sensor is used to measure the atmospheric pressure at a given location. An example of gauge pressure would be a tyre pressure gauge. When the tyre pressure gauge reads 0 PSI, there is 14.7 PSI (atmospheric pressure) in the tyre.It is suggested that it can be used data unstructured data. The data might vary for each pressure sensor of different types.

2.10. Proximity

It is used to detect the nearby object without any physical contact. There are four fundamental proximity sensors [17] they are inductive, capacity, ultrasonic and optical. In inductive proximity the power losses in the coil vary as (metallic).Capacity proximity the internal oscillator does not oscillate until an otherwise the target material is moved. The ultrasonic sensor works if a target object is located in front of the sensor it is applicable only to a certain range. Optical sensor here the sensing material uses the light for sensing so it can sense the large distance. The proximity senor installing guides are represented on the website [18].

3. HOLISTIC VIEW OF DATA IN IOT

From the survey, it is noted that [5] data for the future access will be more on unstructured data. Data can be of two types as we have elaborated earlier. Data can be of two types divided based on business tactics. It is of internal data that consist of CRM, Inventory records, Sales records, sensor data, Online Forum, etc., and another category is external data which contains census data, Twitter, Facebook, etc. The both internal and external can be structured and unstructured. Managing these data represents the business view of data.

The data represents in business must compensate the following aspects:

- Data must be first identified it comes under which category so that storing data quite be much easier. For example, if data is on blogs then it is unstructured data.
- Security of storage if data represents internal data of business value then data must be more secure since it will affect the business development.
- As the data becomes more day-by-day managing huge data is an important aspect. The company has to spend more time in managing data.
- Data retrieved is another difficult task. So only selected data must be retrieved from the database.

4. HANDLING DATA IN IOT

Handling data in IoT is an important task while generating a large amount of data. IoT data storage meet after preprocessing meet the need of data expression form and distribution form. A large volume of heterogeneous data is generated by many sensor devices at real-time. Continuous observation of data is not necessary, but the devices must be of end-to-end response. It will be stored in schema database that change when data in mobile objects change that is the continuous update of data. It consists of various kinds of data types and frequency of data capture for example in humidity we refer data in °C where for chemical sensors we measure in DC so, each sensor generates different types of data that must be equally managed in the database. The relational database is an important structure [20] that stores data in the form of table schema.so that efficient retrieval of data from large storage.

5. VARIETY OF DATA

Big Data is collected from new sources that haven't been mined of insight in the past. Traditional data management processes can't cope with the heterogeneity and variable nature of big data, which comes in formats as different as e-mail, social media, videos, images, blogs, and sensor. First, the unstructured data must be transformed to structured data. Unstructured data also consist of alarm logs, assert data, converting those operations into a structure is a little tedious task. The above chapters are some suggestion to convert it.

6. CONVERTING UNSTRUCTURED DATA TO STRUCTURED DATA SUGGESTIONS

Transforming structured to unstructured is considered in the following norms [13]:

1) Text pattern matching is used to identify large-scale structure. It uses the regular expression such as pattern matching or string matching to match the suitable structure.
2) Transforming the data into table forms with important labels is used to identify common section.
3) Usage of text analytics so that it can be used for further link

Data collected from different sensors are of unstructured form the data is extracted using data extraction one of the primary source technique's used for further processing. Text mining refers to the discovery of new knowledge form a collection of text [21] which is previously unknown. After preprocessing of data, it is processed to text mining.

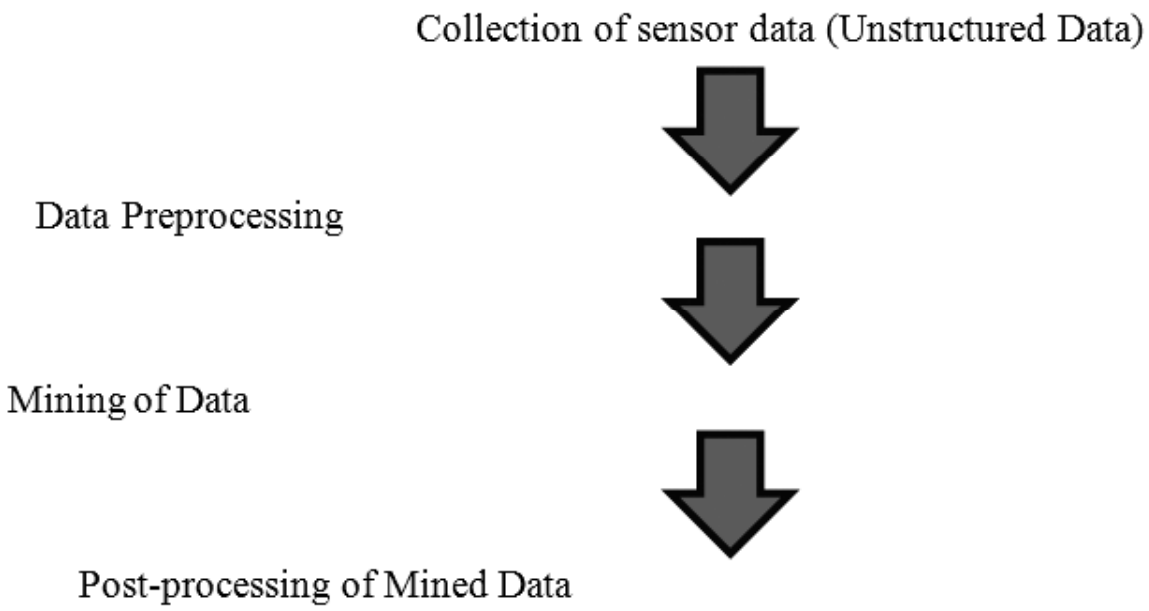

Figure 2: Flow of unstructured data to structured data

The above figure represents the flow from unstructured data to structured data it is found that text is a part of unstructured data in the sensor which is used into consideration.Preprocessing helps to remove noisy, inconsistent data the help to prepare data for further mining. That is from sensors we collected the data and summarized into useful information. Mining is one kind of tool for analyzing data in the received data. In sensors, the collected data is dirty and inconsistent to overcome these preprocessing techniques are applied to both structured and unstructured data. This chapter summarizes some method for data preprocessing. Preprocessing represents [22] cleans the text by removing non-words such as punctuation, special characters. The post preprocessing techniques is identifying useful patterns or interesting patterns.

6.1. Structured Data: Preprocessing Techniques [21] (fixed format)

1. Outliers: Sometimes the data received from sensors are very different from the data that is already present in the database. So those data are eliminated from the database.
2. Missing values: While gathering the data from sensors, there might be the loss of data. The data can be replaced by most frequent appearing data in the sensors
3. Normalization: Process for organizing data in columns and tables.
4. Discretization: Transforming non-categorical attributes to categorical one’s

6.2. Unstructured Data: Preprocessing Techniques [22] (Images, videos ...)

Preprocessing of unstructured data deals with following two problems

1. High Dimensionality of the data: While preprocessing large data might be stored in the database that extends dimension in the database.
2. The sparsity of the data: Distribution of data, some terms occur few, frequent and some may not appear.

Preprocessing Techniques for Unstructured Data–Images [22]

1. Image resampling: Reduce or increase the image pixels to enhance the image appearance.
2. Greyscale contrast enhancement: Improves the visualization of the image.

3. Noise removal: Use filtering techniques such as mean, median, etc., Image pixels value lower than intensity value is removes
4. Mathematical operations: Arithmetic Operation and Morphological Operations are applied to images
5. Manual correction: Editing image to 2D or 3D shapes fine tuning to achieve the clear image.

Preprocessing Techniques for Unstructured Data –Videos [22]

1. Image Cropping and Local Operators: First step in a preprocessing video is cropping that the image is cropped so that it is used to focus on the region for future and reduce the cost.
2. Neighborhood Operators: It is used to remove the noise to get the clear image.
3. Morphology Operators: Here we convert colorful video into binary image for video text detection

7. PROCESSING DATA USING CLUSTERING

7.1. Clustering

7.1.1. k-means Clustering

Algorithm steps for k-means clustering [25]

Let X = Unstructured Data D_1, D_2……D_n (Document), C_m= Cluster centers

Step 1: Randomly choose the cluster centers (Document)

Fun (C_m) =rand (D_1 to D_n)

Step 2: Calculate Euclidean distance

Cal (Dist) = Sqrt(C, D_1,D_2……D_n)

Assign the document to clusters based on minimum distance

Step 3: Calculate centroid for each cluster.Replace with new clusters.

Step 4: Step 2 and Step 3 are repeated until no document moves.

The MATLAB coding for k-means algorithm is available on the website of http://in.mathworks.com/help/stats/kmeans.html?requestedDomain=in.mathworks.com with an example

7.2. Limitation

In K-means algorithm the data are belong to only one cluster and the sensitive to data. Sometimes certain data are the outlier that can influence the current data. They are also sensitive to inner centroids that might be inconsistent.

7.2.1. Fuzzy c-means

It is an extension of K-means here the data may belong to more than one cluster. This algorithm overcomes the drawbacks of K-means algorithm. The MATLAB coding for Fuzzy c-means is available on the website http://in.mathworks.com/help/fuzzy/fcm.html.

Step 1: Place k document into space as objects that are being clustered.

Step 2: Each document is assigned to the closest centroid.

Step 3: After all documents are assigned recalculate the position of the centroid.

Step 4: Repeat step 2 and step 3 until the centroids no longer move

7.3. Limitation

In fuzzy c-means, the iteration is repeated many times for selecting the centroid. Apriori specification of the number of clusters is the main drawback. We move to a Rough k-means algorithm.

7.3.1. Rough k-means

To overcome the inconsistent and incomplete data, the Rough k-means is introduced by Pawlak [27].The above algorithm represents it for every document D rough set is characterized for lower approximation BD and upper approximation B's where B is an indiscernibility relationship

$$BD = \vee\{y \in U / B \,|Y \subseteq D\}$$

$$B'D = \vee \{y \in U / B \,|Y \wedge D \neq \phi\}$$

The modified form of centroids into upper and lower approximation will improve the efficiency of the algorithm and produce better result compared to fuzzy c-means and k-means algorithm.

8. CONCLUSION

In this chapter we have discussed several IoT applications and what kind of data they handle. We have used preprocessing techniques to convert structured data to unstructured data. Our experience shows that it is important to move data to structured form without any data loss and clustering methods provide the major task for this conversion. Among the clustering algorithm, Rough k-means is better to identify useful patterns. This IoT data creates exciting opportunities and challenges to our world.

REFERENCE

[1] Wilson, J.S., 2005. Sensor Technology Handbook. Elsevier, Amsterdam/Boston

[2] Buerk, D. Biosensors. Theory and Applications. Technomic Publishing, Lancaster, 1993.

[3] Niraj, Gupta, "Sensors For Diabetes: Glucose Biosensors By Using Different Newer Techniques: A Review," International Journal of Therapeutic Applications, Volume 6, 2012.

[4] Vijaylakshmi Kamat "Pulse Oximetry" Indian J. Anaesth. 2002; 46

[5] Approaches for Managing and Analyzing Unstructured Data #1N. Veeranjaneyulu, N; Bhat, M Nirupama; Raghunath, AnInternational Journal of Computer Science and Engineering,Jan 2014.

[6] Fan, T. R. and Chen, Y. Z.(2010) A Scheme of Data Management in the Internet of Things.Proceedings of ICNIDC-2010, Beijing,24-26 September 110-114.

[7] A Survey on Gas Sensing Technology,Xiao Liu, Sian Cheng Hong Liu, sensors www.mdpi.com/journal/sensors,2012

[8] Temperature and product model [Online] https://akizukidenshi.com/download/ds/aosong/DHT11.pdf

[9] Power Generation [Online]: http://core.spansion.com/article/energy-harvesting-devices-replace-batteries-in-iot-sensors/#.V4xqfEt97IU

[10] Chemical smoke detector [Online]:http://www.systemsensor.com/en-us/Documents/2151_2151T_DataSheet_A05-0182.pdf

[11] Light Sensor [Online]: http://www.tvdsb.ca/uploads/ScienceProbeware/lightsensor.pdf.

[12] Ambient Light Sensor [Online]:https://msdn.microsoft.com/en-us/library/windows/desktop/dd318933%28v=vs.85%29.aspx.

[13] Data Extraction [Online]: https://en.wikipedia.org/wiki/Data_extraction.

[14] Occupancy Sensor [Online]://www.lutron.com/TechnicalDocumentLibrary/3683197.pdf

[15] Position sensor [Online]:http://sensing.honeywell.com/honeywell-sensing-position-rangeguide-000709-23-en.pdf

[16] Pressure Sensor [Online]:http://www.cynergy3.com/sites/default/files/blog-documents/Explanation%20of%20pressure%20sensors.pdf

[17] Proximity Sensor [Online]:http://staff.iha.dk/jse/Elteknik%20noter/Sensorer/8-3%20Proximity%20Sensors.pdf

[18] Technical guide of proximity sensor [Online]: https://www.ia.omron.com/data_pdf/guide/41/proximity_tg_e_6_1_2%28classifications%29.pdf

[19] SapnaTyagi,Ashraf Darwish,Mohammad [2014],Managing Computing Infrastructure for IoT data ,Advances in Internet of Things.

[20] Ramakrishnan, R.; Gehrke, J. Database Management Systems, 3rd ed.; McGraw-Hill: New York, NY, USA, 2002.

[21] Santosh Kumar Paul, Madhup Agrawal, Shyam,(2014) An Information Retrieval(IR) Techniques for Text Mining on a web for Unstructured data, Advanced Developments in Engineering and Technology.

[22] Martin, Andreas, (2016) Data Preparation for Big Data Analytics: Methods and Experience, Enterprise Big Data Engineering Analytics and Management, IGI Global.

[23] Preprocessing for Images [Online]:https://www1.imperial.ac.uk/resources/460110AC-29CD-49DE-A9BF-E4F7ED301430/

[24] Video Preprocessing, Chapter 2, Advances in Computer Vision and Pattern Recognition, Springer

[25] Evolving limitations in K-means algorithm in data mining and their removal,[2011], Kehar Singh, Dimple Malik and Naveen Sharma, IJCEM International Journal of Computational Engineering & Management, Vol. 12,

[26] An Improved Algorithm of Rough K-Means Clustering Based on Variable Weighted Distance Measure[2014], Tengfei Zhang, Long Chen and Fumin, International Journal of Database Theory and Application, Vol.7, No.6,pp.163-174