ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

Data Mining Algorithms for Internet of Things: Technical Review

Vikas B O[*1], Dr.Jitendranath Mungara[*2]

Asst. Prof, Department of Information Science Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India*[1]

HOD, Department of Information Science Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India[2]

ABSTRACT: In today's world, everything is connected via internet, but Internet of Things will change our life in the future. For, this to happen, large amount of data has to be generated, processed and captured by IoT are considered to have highly useful and valuable information. The critical role to be played in making things smart is through Data Mining Methodology. This paper focus on systematic review on data mining concepts such as generating data, processing data and capturing data and "Data Mining algorithms for IoT". Finally, a suggested big data mining system is proposed along with the potentials, challenges, open/closed issues and future trends of data mining in IoT fields.

KEYWORDS: Data Mining, Internet of Things

I. INTRODUCTION

The Internet of Things (IoT) and its has been considered the technology that combines networks, devices, instrument and objects. Connecting all the objects [1], forming a network of devices is the basic idea of IoT. The current world that is connected to the internet, IoT uses the same internet to connected devices that can be easily monitored and controlled, also the same things can be automatically detected by other things, further communicate with each other through internet, and can even make decisions themselves [2]. In order to make this decision more accurate, analysis of the data generated has to be captured efficiently and processing the same requires technologies. One of the most extremely useful technologies is Data Mining.

The large amount of data being generated from IoT devices is stored in datasets. From these large datasets, extraction of useful information i.e., discovering useful patterns to provide efficient extraction of hidden information. The large amount of data that is processed with data mining methodology predicts the model, also generalize to new data. Thus, Data mining is the process of discovering interesting knowledge from large amounts of data stored in repositories, data warehouses or the database.

Typical stages involved in data mining process is as follows:

1. Data Construction
2. Data Mining
3. Data Dispensing.

Data Construction: Constructing the large data for mining. This stage involves four substeps: Combine data from various origin and noise cleansing of data; data mining system must be given input of some parts of data; finally process the data to the data mining stage.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

Data Mining: Algorithm to find the patterns and evaluating patterns from the discovered information. **Data Dispensing:** Mined information and visualizing the data to the end user.

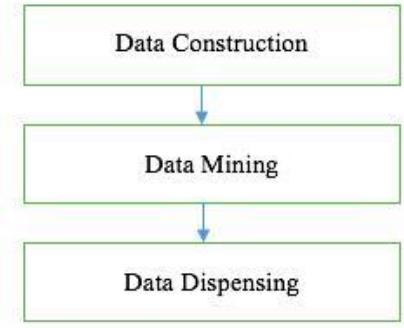

Fig 1: Data Mining Stages

II. DATA MINING FUNCTIONALITIES

The data mining functionalities include the following levels:

a. Classification Study

b. Clustering Study

c. Association Study

d. Outlier Study

e. Time Series Study

a) **Classification Study** is a data mining function that assigns items in a collection to target categories or classes.
b) **Clustering Study** finds clusters of data objects that are similar to each other and form a group called cluster.
c) **Association Study** is a data mining function that finds the probability of the occurrences of an items in a collection of data information
d) **Outlier Study** is an observation point that is distant from other observations

e) **Time Series Study** is a series of data points listed in time order.

2.1 Classification Study

In decision making classification plays an important role for managing of large data. When the data is given, it is assigns to one of the target class or category called classification [3]. Accurately predicting towards the target category is the goal of classification. For example: Low, Medium or High classes. [4]

Various methods in classification are as follows:
a. Decision Tree
b. KNN
c. Bayesian Network

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

d. SVM

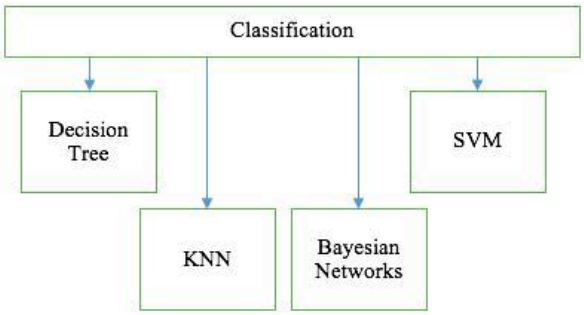

Fig 2: Classification Structure

The suitable classification algorithms for IoT devices input RAW data to provide efficient results can be:

a. KNN Classification

b. Naïve Bayes Classifier

2.1.1 K-Nearest Neighbors Classification

K Nearest Neighbors [5] is an algorithm that stores all the existing data into cases, further use the same to classify data into new cases based on similar measure i.e., distance based functions.

Example: In this example, a shopping store has products/things that are connected over the internet say IoT connected. Here, based on the customers previous shopping details/cases, classification is performed to provide suggestion/recommendations for the similar products through k-NN algorithm.

2.1.1.1 Training set: (x1, y1), (x2,y2),…, (xn,yn)

Initially training data from previous purchase history of customer cases are stored. This data is used to classify and determine the accuracy of k-NN algorithm.

Table 1: Data set for training data

X1= Product Type	X2= Product Size	Y= Classification
3	5	**Reject**
7	5	**Reject**
5	5	**Suggest**
8	4	**Suggest**

Consider x1=3 and x2=7 and K=3

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

2.1.1.2 Distance Calculation

Calculate the distance between new cases with all the existing training data.

Table 2: Distance Calculation

Distance Calculator: $(X1-x1)^2 + (X2-x2)^2$
$(3-3)^2+(5-7)^2= 4$
$(7-3)^2+(5-7)^2=12$
$(5-3)^2+(5-7)^2=8$
$(8-3)^2+(4-7)^2=34$

2.1.1.3 Nearest Neighbor discovery

With the input from table 2, sort the distance to discover the nearest neighbour based on the kth minimum distance.

Table 3: Nearest Neighbor Sorting

Calculate Distance	**Minimum Distance Rank**	**Is it included in 3 nearest neighbour (k=3)**
4	1	Yes
12	3	Yes
8	2	Yes
34	5	No

2.1.1.4 Finding the category Y classification of the nearest neighbor

Table 4: Category Y of the nearest neighbour

X1= Product Type	**X2= Product Size**	**Distance Calculation**	**Rank Of Min. Distance**	**k=3 nearest neighbour**	**Y**
3	5	4	1	Yes	**Reject**
7	5	12	3	Yes	**Reject**
5	5	8	2	Yes	**Suggest**
8	5	34	5	No	-

With the results obtained from the above table 4, the classification shows 2 reject and 1 suggest.

Predictions in k-NN is based on majority of votes and since there are two votes for reject and one vote for suggest. Therefore, it concludes that the product is rejected.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

2.1.2 Naïve Bayes Classifier

The aim of Naïve Bayes classifier [6] is to construct a rule that performs addition of new/future objects to the class. In order to demonstrate/use this to produce classifications, we need the help of training dataset and early estimation on the same. Once all the possible probability from a given dataset is generated, if any new/future data arrives, calculation of probability with training dataset provides end results indicating to which class the input data belongs i.e., classification is performed.

Example: Shopping store example where customer receives recommendation on product reject/suggest.

2.1.2.1 D: Set of tuples

Each tuple is an 'n' dimensional attribute vector X: (x1,x2,x3,…., xn)

X=(Product type = 5 , Product Size = 5)

Table 1: Dataset.

Product Type	Product Size	Buys Product
3	5	Yes
7	5	Yes
5	5	Yes
8	4	No

2.1.2.2 Probability determination for each class

Probability for feature vector is

X= (product type=5, product size=5)

P(X/Ci) = P(x1/Ci) * P(x2/Ci) * . . . . . P(xn/Ci)

P(c1) = (P(C1) = P (buys product = Yes) = 3/4

P(c2) = (P(C2) = P(does not buy product = No) = 1/4

P(Type= 5 / Buy Product = yes) = 1/3 = 0.33

P(Type= 5 / Buy Product = no) = 1/1 = 1

P(Size= 5 / Buy Product = yes) = 1/3 = 0.33

P(Size= 5 / Buy Product = no) = 1/1 = 1

P(X/buys product = Yes) = P(product type=5/buys product=yes) * P(product size=5/buys product=yes) = 0.33

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

P(X/buys products = No) = 0.33

2.1.2.3 Find class Ci

P(X/Ci) * P(Ci)

P(X/buys product= yes) * P(buys product= yes)=0.24 P(X buys product= no) * P(buys product= no)=0.08

Since the probability of buying the product is 0.24 that is greater than of not buying 0.08. This shows that customer X will buy the product.

2.2 Clustering Study

Clustering does dividing the data into meaningful sets/groups[55]. The groups are formed with similar data patterns into one and dissimilar data patterns into another. In case of large data, clustering extracts useful information and creates groups called clusters. This is also termed as unsupervised learning [4].

Various methods in clustering are as follows:

a. Partitioning

b. Hierarchical

c. Co-occurrence

d. High Dimensional

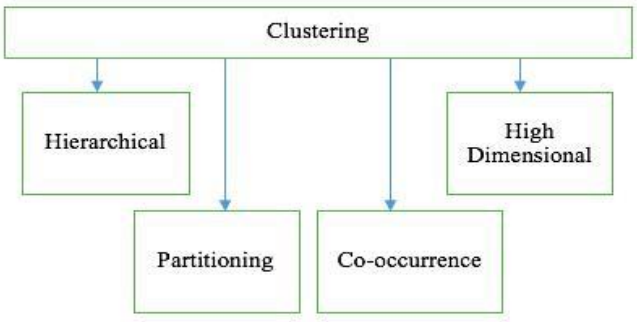

Fig 3: Clustering Structure **a. Partitioning**

k-means clustering algorithm [7] is one of the simplest unsupervised algorithm. It follows an easy way to classify a given data set through a certain number of clusters.

Example: Shopping store example where customer tries to purchase an item in store at certain location and the item is available in another store. Here the retailer has to track customer behaviour by sensors embedded in products to provide contextual information. The historical data can be analysed (i.e., customer recently visiting store location, products etc) and allow retailer to recommend certain products.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

In order to perform the above recommendation, we need analyse the data information from various store location, cluster them as groups and can be used to send otifications/alerts/recommendation to the customer, we need to mine the data using **k-Means Clustering algorithm.**

Algorithm:

Let X= { x1,x2,x3,…,xn} be the set of products and V = {v1, v2, v3,…,vn} be the store locations(clusters). **Step 1:** Randomly select c cluster centres

Step 2: Calculate distance between each product and store locations.
Step 3: Assign the product item to the store (cluster) location whose distance is minimum of all the store locations.

Step 4: Recalculate new store location (cluster) using:

$$\mathbf{Vi= (1/Ci) \Sigma Xi}$$

where, 'Ci' represents number of products in i^{th} store location(cluster).

Step 5: Recalculate the distance between each product data point and new obtained store location (cluster) centers.

Step 6: If no products were reassigned then stop, otherwise repeat from step 3.

Working of algorithm:

Given X={2, 4, 10, 12, 3, 20, 30, 11, 25} are products, assume store k=2.
Based on the user historical data, we can cluster the products to the store locations where customer usually buys the products,

k1= {2, 10, 3, 30, 15} = 14

k2= {4, 12, 20, 11} = 11.75

Re-assign the store location (cluster), by selecting the products closer to mean value (history data).
k1= { 2, 3, 4, 11, 12 } = 7

k2= {20, 30, 15 } = 25
Repeat the steps again,
k1= { 2, 3, 4, 11, 12 } = 7

k2= {20, 30, 15 } = 25

Finally, store location_1 (cluster1) will have products 2, 3, 4, 11, 12 and store location_2 (cluster2) will have products 20,30,15.

2.3 Association Study

Market basket analysis and transaction data analysis is the methodology of Association rule mining [8], that produces some set of rules showing attributes-value associations that occur frequently. This helps in the generation of more general and qualitative knowledge, which in turn helps in decision-making [9].

Various methods in association are as follows:

a. Parallel

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

b. Temporal Sequence
c. Priority/Pattern Sequence
d. Genetic and Fuzzy Set

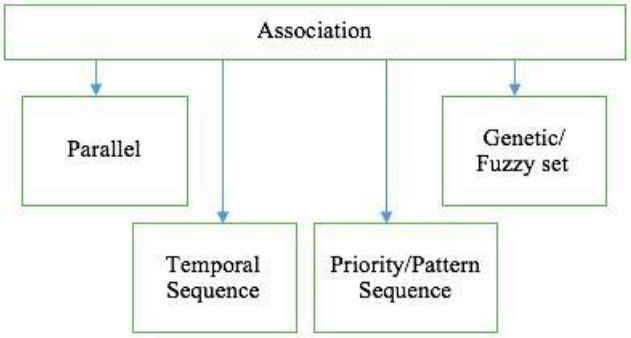

Fig 4: Association Structure

Example: Shopping store example where customer tries to purchases a product, recommendation of frequently purchased products in the same store associated with current product. This recommendation can be done by retailer using association rule mining algorithm named

apriori algorithm.
Working of algorithm:
Given customers product history transactions

Table1: Customer Transaction History

Transaction ID	Products Brought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

Step 1: Scan product history transaction table for each products { 3, 4, 5, 6, 7, 8 }and find the support P1,

Table 2: Products with Support

Products	Support
{ A }	3
{ B }	2
{ C }	2
{ D }	1
{ E }	1
{ F }	1

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

Step 2: Compare the product support count with minimum support count (say 50%)L1=

Products	Support
{ A }	3
{ B }	2
{ C }	2

Step 3: Generate product P2 from L1 and scan for count of each product P2 and find the support C2 =

Products	Support
{ A, B}	1
{ A, C }	2
{ B, C }	1

Step 4: Compare product (P2) support count with minimum support count L2 = { A, C } = 2
Step 5: Therefore association rule that can be generated from L are as shown below with support and confidence.

Association Rule	Support	Confidence	Confidence %
A->C	2	2/3=0.66	66%
C->A	2	2/2=1	100%

As minimum confidence threshold is 50% (as set by retailer). With these rules, product A and product C becomes the frequently purchased products and the same can be recommended for the customer.

2.4 Outlier Study

The method of finding patterns in data that are different from the rest of the data with the help of appropriate metrics is called Outlier detection. Sometimes, the abnormal behaviour of the system described by the data becomes useful information pattern. This is obtained by distance based algorithms that calculates distances among objects in the data with interpretation geometrically. [10]

2.5 Time Series Study

A collection of temporal data objects is the time series study. This includes a characteristic namely data size, data dimensionality along with updates. [11]

Various components in time series study are:

a. Representations

b. Similarity Measures

c. Indexing

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

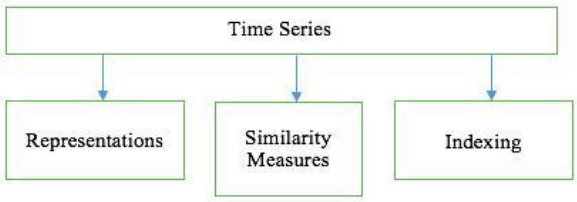

Fig 5: Time Series Structure

III. CHALLENGES AND OPEN ISSUES IN IOT

The most fundamental issue in IoT, big data and cloud computing with its enormous development is how to process the large volumes of data and extract the useful information that can predict future actions[12]. Various key elements that must be considered in IoT era are as follows:

(i) Read and Write of Large volumes of data: The need of effective computation mechanism should be identified in order to process the data that can be in TB (Tera Bytes), PB (Peta Bytes) and ZB (Zetta Bytes).

(ii) Integration of data types to Heterogeneous data sources: In IoT devices, the need of integration of camera data, sensors data [13] and online social networks data and so on. All these data will be in different formats, hence the need of communication between data types such as numbers type, binary type, bytes type and so forth. Also, communication between different types of devices, systems and an efficient mechanism to extract data from these IoT devices is at high priority.

(iii) Extraction of Complex Data:

In large volume of data, extraction of useful information i.e., hidden data extraction is required in order to evaluate and analyse the different properties of data and associate the same to provide the result is required.

3.1 Challenges in IoT

As the quantity of data is in large volumes in IoT and data is received from various data sources that has different types, formats and representation forms, also the data is heterogeneous, structured, semi-structured an sometimes unstructured happened when IoT devices fetch the data from its various parts. This shows that various challenges exists when it comes to IoT and Big data.

Challenges can be categorised into two stages:

(i) First Challenge: The process of generating, accessing, capturing and extracting data from different data sources that deal with data having heterogeneity, noise, variety and faults. The upmost challenge is to find the fault

and rectify the fault in order to correct the data and perform processing. This is a biggest challenge for data mining algorithms i.e., how to modify traditional algorithms to big data environment.

(ii) Second Challenge: The mining of incomplete, faulty and uncertain data for large volumes of data is the biggest challenge. An efficient solution for sharing data between different systems and applications (sensors, cameras etc.) with security is one of the challenge, as

sensitive information such as medical records, home-security systems, banking transactions is a matter of concern.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

IV. SUGGESTED ARCHITECTURE MODEL FOR IOT

This architecture model is designed for IoT system and big data mining system.

It consists of three core layers and sub layers as follows:

a. Devices: IoT devices such as RFID, Camera, Sensors and other devices can be integrated for generating the data.
b. RAW data: Structured, unstructured, semi-structured data can be integrated from the RAW data from IoT devices.
c. Data Collection: Data can be collected from various forms such as real-time data, batch data, parsed data, analyzed data and merged data.
d. Data Processing: Open source solutions can be used at this layer such as HDFS, Hadoop, Spark.
e. Service Registry: Mining functions can be kept at service registry
f. PSS- Privacy/Standard/Secure: In Data mining system these three components play vital role in order to secure data from unauthorized access.

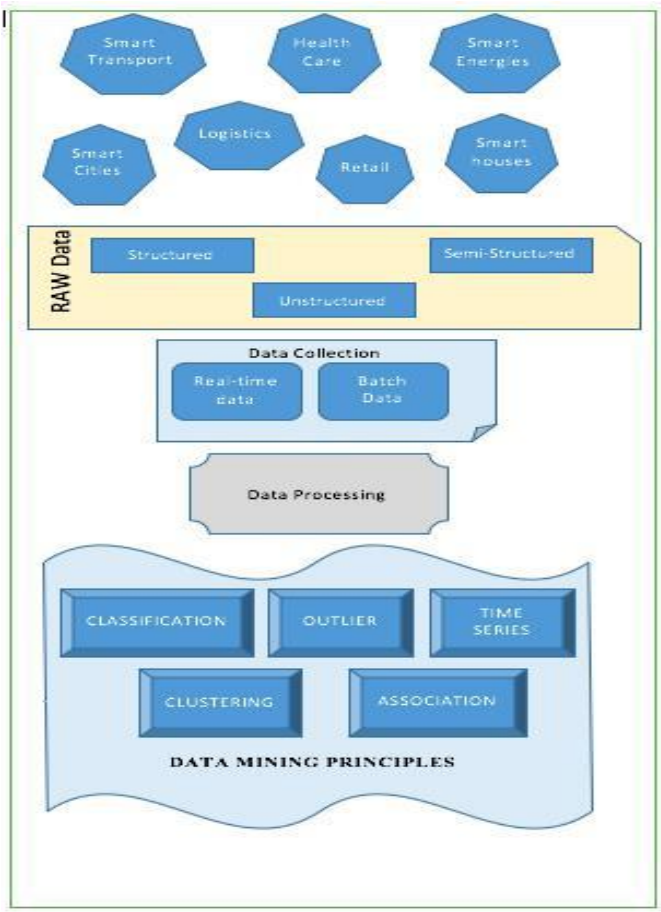

Fig 6: Architecture Model for IoT

V. CONCLUSION

The need to manage the large volumes of data, automate the processes via exploring devices and its connectivity through sensors in IoT environment is at high priority. In order to provide decisions for both end user and IoT devices the integration of data mining methodology is required that perform decision making support and optimization of systems. Extraction of useful information from large data can be done only by applying algorithms with patterns,

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 5, June 2017
8th One Day National Conference on Innovation and Research in Information Technology (IRIT- 2017)
Organized by
Departments of ISE, CSE & MCA, New Horizon College of Engineering, Bengaluru, Karnataka 560103, India

interestingness and knowledge discovery for efficient analysis. Nowadays, IoT devices are increasing and producing lots of data over the internet, thus leading to big challenges on how to analyse the data and how to mine the data effectively. Based on the survey, a suggested architectural model for IoT is proposed.

REFERENCES

[1] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the internet of things: perspectives and challenges," Wireless Networks, vol. 20, no. 8, pp. 2481–2501, 2014.
[2] C.-W. Tsai, C.-F. Lai, and A. V. Vasilakos, "Future internet of things: open issues and challenges," Wireless Networks, vol. 20, no. 8, pp. 2201–2217, 2014.
[3] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT '13), pp. 1–7, July 2013.
[4] S. Song, Analysis and acceleration of data mining algorithms on high performance reconfigurable computing platforms [Ph.D. thesis], Iowa State University, 2011.
[5] D. T. Larose, "k-nearest neighbor algorithm," in Discovering Knowledge in Data: An Introduction to DataMining, pp. 90–106, JohnWiley & Sons, 2005.
[6] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence, pp. 399– 406, 1994.
[7] Q. Li, P. Wang,W.Wang, H. Hu, Z. Li, and J. Li, "An efficient K-means clustering algorithm on MapReduce," in Proceedings of the 19th International Conference on Database Systems for Advanced Applications (DASFAA '14), Bali, Indonesia, April 2014, vol. 8421 of Lecture Notes in Computer Science, pp. 357– 371, Springer International Publishing, 2014.
[8] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the ACMSIGMOD, International Conference on Management of Data (SIGMOD '93), pp. 207–216, 1993.
[9] A. Gosain andM. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining," in Proceedings of the IEEE Conference on Information & Communication Technologies (ICT '13), pp. 1003–1008, JeJu Island, Republic of Korea, April 2013.
[10] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," The Computer Journal, vol. 54, no. 4, pp. 570– 588, 2011.
[11] T.-C. Fu, "A review on time series data mining," Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 164–181, 2011.
[12] T. Hu, H. Chen, L. Huang, and X. Zhu, "A survey of mass data mining based on cloud-computing," in Proceedings of the International Conference on Anti-Counterfeiting, Security and Identification (ASID '12), pp. 1–4,August 2012.
[13] Y. Sun, J. Han, X. Yan, and P. S. Yu, "Mining knowledge from interconnected data: a heterogeneous information network analysis approach," in Proceedings of the VLDB Endowment, pp.2022-2023, 2012.

BIOGRAPHY

VIKAS B O is an Assistant Professor in the department of Information science and engineering, New Horizon College of Engineering, Bangalore, India. He has received his B.E. degree in Computer Science and Engineering in 2013 and M.Tech degree in Computer Science and Engineering in 2015. His research interest includes Data mining, Information security and Internet of things. He is a member of Computer society of India, Mumbai, India. He has received two best paper awards in international conferences.

Dr. Jitendranath Mungara is a Double Doctorate in Computer Science and System Engineering and Electronics. He is working as a Prof. & HOD of Information Science and Engineering department in NHCE Bangalore. He is pioneering in data mining and IoT core research. He has published 75 papers in International Journals and Conference. Obtained three doctorates under his guidance and currently guiding three Ph.D students in MANETS and IoT.