MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 7, Issue 2, July-December 2022

Using Data Mining Techniques for Internet of Things

Aakash Nag[1], Sumit Porwal[2], Dr. Ashish Chaturvedi[3]

Assistant Professor[1], Students[2, 3]

Department of Computer Science and Engineering

Vishwakarma Sahajeevan Madanbhai Sura Institute of Engineering & Technology

***Corresponding Author's Email:** -aakashnag745@gmail.com[1]*

Abstract

A smart world, such as ours, is essentially built on the Internet of Things (IoT). Every day, the IoT creates massive amounts of data. It is critical to work on this generated data in order to make appropriate use of it. This is mostly accomplished through the use of data mining. This research examines how data mining may be used on IoT data. Classification (grouping data), clustering (labelling data), frequent pattern mining (identifying often occurring itemsets, sequences, or substructures in data), and outlier analysis are the main components of data mining (analyzing data with abnormal values of attributes). This paper's major focus is on frequent pattern mining.

***Keywords**: Frequent Pattern Mining, Internet of Things, Data Mining*

INTRODUCTION

The Internet of Things (IoT) is a network of physical devices, such as automobiles, household appliances, and other items that are integrated with electronics, software, sensors, actuators, and connections, allowing these objects to communicate and share data. Each object is uniquely identified by its embedded computing system, yet it may interact with the existing Internet infrastructure. [1, 2, 3]

From 2016 to 2017, the number of online-capable gadgets climbed 31% to 8.4 billion. [4] Experts predict that the Internet of Things will have 30 billion items by 2020 [5]. The worldwide market value of the IoT is expected to reach $7.1 trillion by 2020. [4]

The Internet of Things enables objects to be sensed or controlled remotely across existing network infrastructure, allowing

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

for more direct integration of the physical world into computer-based systems and resulting in improved efficiency, accuracy, and economic benefit, as well as reduced human intervention. When sensors and actuators are added to IoT, it becomes a subset of the larger class of cyber-physical systems, which also includes smart grids, virtual power plants, smart homes, intelligent transportation, and smart cities.

Things in the context of the Internet of Things can refer to a wide range of devices, including heart monitoring implants, biochip transponders on farm animals, cameras streaming live feeds of wild animals in coastal waters, automobiles with built-in sensors, DNA analysis devices for environmental/food/pathogen monitoring, and field operation devices that assist firefighters in search and rescue operations. According to legal academics, "things" are a "inextricable blend of hardware, software, data, and service." [5] These gadgets capture important data using existing technologies and then autonomously transfer the data to other devices. Figure 1 depicts an IoT technological roadmap, demonstrating how IoT has evolved over time and what the future holds.

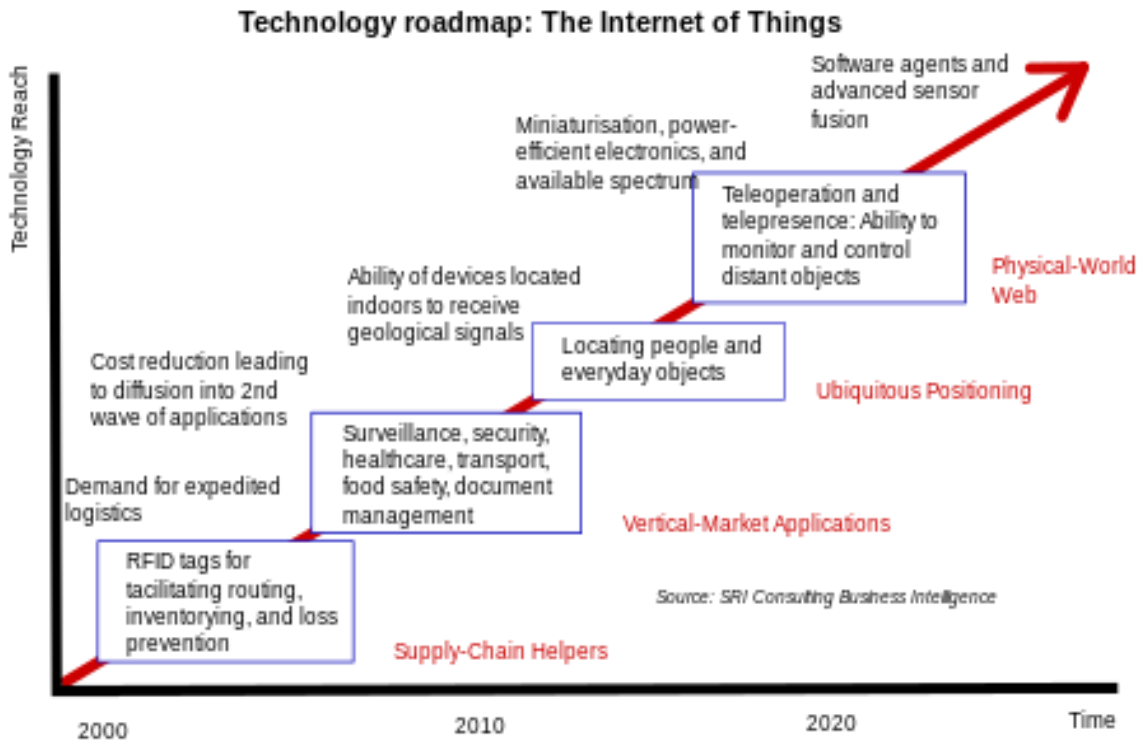


Figure 1: Technology roadmap for IoT

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Varying studies [6, 7, 8, 9] estimate different quantities of data created globally each year; nonetheless, it is assumed that the overall quantity of data generated has recently topped one zettabyte. It is obvious that today's data analysis tools are just not powerful enough to process and evaluate large amounts of data from IoT. There is little question that fitting more than one zettabyte into a single storage device remains a tough task. The second point to consider is that IoT data is typically too large and difficult to handle with the technologies available today. According to Baraniuk [6,] the bottleneck of data processing will be relocated from the sensor to the sensor's data processing, transmission, and storage capability.

Data mining is the process of detecting patterns in huge data sets using methods from machine learning, statistics, and database systems. [10] It is a necessary procedure in which intelligent approaches are used to extract patterns from data. Data mining tasks are divided into six categories [10]:

1. Frequent pattern mining - Looks for correlations between variables. A supermarket, for example, may collect information on client purchase behaviour. The store can employ frequent pattern mining to identify which goods are commonly purchased together and use this knowledge for marketing purposes. This is also known as "market basket analysis."

2. Clustering: the problem of identifying groupings and structures in data that are "similar" in some manner without using known structures in the data

3. Classification is the job of applying existing structure to new data. E-mail software, for example, may attempt to categorise an e-mail as "legitimate" or "spam."

4. Outlier detection: Identifying unexpected data records that may be intriguing or data mistakes that need to be investigated further.

5. Regression seeks to discover a function that models the data with the least amount of error, in order to estimate the correlations between data or datasets.

6. Summarization entails giving a more compact representation of the data collection, as well as visualisation and report production.

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

This paper's major focus is on frequent pattern mining and how it may be used to mine data, specifically IoT data.

FREQUENT PATTERN MINING ON IOT DATA

Pattern mining is the process of discovering intriguing, surprising, and beneficial patterns in databases by using or creating data mining algorithms. Pattern mining methods may be used on a wide range of data types, including transaction databases, sequence databases, streams, texts, geographical data, graphs, and so on. Sub-graphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns, high-utility patterns, and so on can be discovered using pattern mining algorithms.

Apriori is the most widely used pattern mining algorithm. [11] It is intended to be used on a transaction database to detect patterns in client transactions in retailers. However, it may also be used in a variety of other applications, including IoT data. The sole requirement is that the data be in transactional data format. A transaction is defined as a collection of separate elements (symbols). Apriori accepts as input (1) a user-specified minsup (minimum support) threshold and (2) a transaction database comprising a collection of transactions. Apriori returns all frequent itemsets, which are groupings of items that are shared by at least three minor transactions in the input database.

The Apriori method uses an iterative level-wise technique based on candidate generation to locate common itemsets. The following is a summary of the procedure:

Input:

D, a database of transactions;

min sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

L_1 = find frequent_1-itemsets(D);

for ($k = 2$; $L_{k-1} \neq \Phi$; k++)

C_k = apriori_gen(L_{k-1});

for each transaction $t \in D$ { // scan D for counts

C_t = subset(C_k, t); // get the subsets of t that are candidates

for each candidate $c \in C_t$

c.count++;

$L_k = \{c \in C_k | \text{ c.count} \geq \text{min_sup}\}$

}

return $L = U_k L_k$;

procedure *apriori_gen*(L_{k-1}: frequent(k-1)-itemsets)

for each itemset $l_1 \in L_{k-1}$

for each itemset $l_2 \in L_{k-1}$

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

if ($l_1[1] = l_2[1]$ ^ ($l_1[2] = l_2[2]$) ^ ... ^ ($l_1[k-2] = l_2[k-2]$ ^ ($l_1[k-1] < l_2[k-1]$) then

{

$c = l_1$ *join* l_2; // join step: generate candidates

if has_infrequent_subset(c, L_{k-1}) then

delete c; // prune step: remove unfruitful candidate

else add c to C_k;

}

return C_k;

procedure *has_infrequent_subset*(c: candidate k-itemset; L_{k-1}: frequent (k-1)-itemsets)

for each (k-1)-subset s of c

if s ∉ L_{k-1} then return TRUE;

return FALSE;

Algorithm : Apriori

I executed Apriori on a transactional dataset (see Table 1). It contains four transactions. Given a minsup of 2 transactions, frequent itemsets are “bread, butter”, “bread milk”, “bread”, “butter” and “milk” (Table 2).

Table 1: Transactional database

Transaction_ID	Items_in_the_Transaction
T100	bread, butter, coffee
T101	butter, tea
T102	bread, milk, butter
T103	candy, bread, milk

Table 2: Frequent itemsets and their support

Frequent_Itemsets	Support
Bread	3
Butter	3
milk	2
bread, butter	2
bread, milk	2

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Another extended use of this approach is that it can be used for finding association rules. A simple example that is often used to explain the concept of association rule is discovering items that will be purchased together with items already purchased. Mathematically, the problem can be defined as follows: Given a set of items $I = \{i_1, i_2, \ldots, i_m\}$ and a set of transactions $D = (t_1, t_2, \ldots, t_n)$ where $t_i \subseteq I$, find the set of association rules which are greater than or equal to the predefined threshold values of support and confidence. In other words, the two important conditions to evaluate the mining results, support and confidence, are predefined by the user. For example, a transaction rule for buying bread and milk together, denoted {bread} ⇒ {milk}, with the value of support set equal to 10% and the value of confidence set equal to 70%, indicates that 10% of the customers buy bread and milk together while each customer has a 70% of chance to buy milk if he or she bought bread. Mathematically, the support is defined as

$$\text{support}(X \Rightarrow Y) = P(X \cup Y) = [TC(X \cup Y)]/n \quad (1)$$

and the confidence is defined as

$$\text{confidence}(X \Rightarrow Y) = P(Y \mid X) = [TC(X \cup Y) / TC(X)] \quad (2)$$

where TC(a) denotes the number of transactions in D that contain a.

Analyzing purchasing behaviour, like other ways of the association rule, continues to pique the interest of researchers and firms employing RFID or even IoT technologies. Schwenke et al. presented a high-level framework for describing agent states and client behaviour in a supermarket, specifically, thinking, moving, and acting [12]. To address the issue of customers not being able to find the products they are looking for quickly, [13] integrates customer-specific rules (e.g., age and family state), category-based rules, and association rules into the same buying suggestion system to assist supermarket customers in finding the products they are looking for. The rule-based and case-based reasoning are used to evaluate data in order for the system to more correctly advise consumers on what they need based on (1) the personal preference of a customer, (2) their personal purchase history, and (3) the behaviour of other individuals.

The Smart environment is another technique based on RFID and sensor technologies that requires mining technologies to make it more intelligent in order to provide more convenient services.

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Unlike most classification studies, which use activities of daily living (ADL) to characterise the activities, a collection of movements is used to help the classification algorithm discriminate between distinct occurrences. The relationships between temporal activities are the topic of [14, 15, 16]. Several research [14] were inspired by [17] and focused on characterising and identifying the relationships of temporal activities. Before, after, meets, met-by, overlaps, overlapped-by, starts, started-by, finishes, finished-by, during, contains, and equals are used to describe the relationships between events in time.

The apriori technique is also used to detect recurrent patterns in [14]. Later research [18, 19] uses the k-means clustering approach to categorise the activities in order to generate a normal mixture model for each activity before using association rules mining. Another study [20] combines association rule mining with the linear support vector machine (LSVM) classification method to increase the accuracy of a behaviour prediction system for home health care. To recognise the events, patterns acquired by four distinct types of sensors—the ECG sensor, temperature sensor, network camera for human position and mobility, and face expression sensor—are employed. The LSVM is utilised to recognise home services, while the association rule analyses home services for the human if any unexpected occurrences occur, such as a human suffering from stress.

I. CONCLUSION

A technical roadmap for the IoT was outlined in this study, followed by a discussion of which data mining techniques may be employed on IoT data. Then, it was demonstrated how to do frequent pattern mining on IoT data in the form of a transactional database. Finally, association rule mining was presented and examined in terms of how it may be used for IoT data.

REFERENCES

1. C. Schwenke, V. Vasyutynskyy, and K. Kabitzsch, “Simulation and analysis of buying behavior in supermarkets,” in Proc. IEEE International Conference on Emerging Technologies and Factory Automation, 2010, pp. 1–4.

2. M. Hilbert and P. L´opez, “The world’s technological capacity to store, communicate, and compute

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

information," Science, vol. 332, no. 6025, pp. 60–65, 2011.

3. E. Nazerfard, P. Rashidi, and D. Cook, "Discovering temporal features and relations of activity patterns," in Proc. IEEE International Conference on Data Mining Workshops, 2010, pp. 1069–1075.

4. J. Choi, D. Shin, and D. Shin, "Ubiquitous intelligent sensing system for a smart home," in Proc. International Conference on Structural, Syntactic, and Statistical Pattern Recognition, 2006, pp. 322–330

5. J. Gantz and D. Reinsel, "Extracting value from chaos," 2011, IDC IVEW, Available at http://www.itu.dk/people/rkva/2011-Fall SMA/readings/ExtractingValuefromChaos.pdf

6. Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann

7. Noto La Diega, Guido; Walden, Ian (1 February 2016). "Contracting for the 'Internet of Things': Looking into the Nest". Queen Mary School of Law Legal Studies Research Paper No. 219/2016

8. E. Nazerfard, P. Rashidi, and D. J. Cook, "Using association rule mining to discover temporal relations of daily activities," in Proc. International Conference on Toward Useful Services for Elderly and People with Disabilities: Smart Homes and Health Telematics, 2011, pp. 49–56.

9. G. Papamatthaiakis, G. C. Polyzos, and G. Xylomenos, "Monitoring and modeling simple everyday activities of the elderly at home," in Proc. IEEE Conference on Consumer Communications and Networking Conference, 2010, pp. 617–621.

10. R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. ACM SIGMOD International Conference

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved

on Management of Data, vol. 22, no. 2, 1993, pp. 207–216.

11. E. Nazerfard, P. Rashidi, and D. Cook, “Discovering temporal features and relations of activity patterns,” in Proc. IEEE International Conference on Data Mining Workshops, 2010, pp. 1069–1075.

12. Hsu, Chin-Lung; Lin, Judy Chuan-Chuan (2016). "An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives". Computers in Human Behavior. 62: 516–527. doi:10.1016/j.chb.2016.04.023.

13. D. Reed, D. Gannon, and J. Larus, “Imagining the future: Thoughts on computing,” Computer, vol. 45, no. 1, pp. 25–30, 2012.

Page 65- 73 © MANTECH PUBLICATIONS 2022. All Rights Reserved