ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# MINING INTERNET OF THINGS DATA

**Mudasir Shafi[*1], Sumiran Daiya[2] and Sumit Dalal[3]**
[*1,2&3]Sat Kabir Institute of Technology and Management, Jhajjar

**DOI**: 10.5281/zenodo.1247118

## ABSTRACT
A smart world, like ours, is primarily based on the concept of Internet of Things (IoT). IoT generates huge amount on data on a daily basis. It is very important to work on this generated data so as make good use of it. Data mining is essentially used for this purpose. This paper discusses how data mining can be implemented on IoT data. Data mining primarily includes classification (grouping data), clustering (labeling data), frequent pattern mining (finding frequently occurring itemsets or sequences or substructures in data) and outlier analysis (analyzing data with abnormal value of attributes). The main focus of this paper is frequent pattern mining.

**Keywords:** Internet of things, data mining, frequent pattern mining.

## I. INTRODUCTION
Internet of Things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and connectivity which enables these objects to connect and exchange data. Each thing is uniquely identifiable through its embedded computing system but is able to inter-operate within the existing Internet infrastructure. [1, 2, 3]

The figure of online capable devices increased 31% from 2016 to 8.4 billion in 2017. [4] Experts estimate that the IoT will consist of about 30 billion objects by 2020 [5]. It is also estimated that the global market value of IoT will reach $7.1 trillion by 2020. [4]

The IoT allows objects to be sensed or controlled remotely across existing network infrastructure, creating opportunities for more direct integration of the physical world into computer-based systems, and resulting in improved efficiency, accuracy and economic benefit in addition to reduced human intervention. When IoT is augmented with sensors and actuators, the technology becomes an instance of the more general class of cyber-physical systems, which also encompasses technologies such as smart grids, virtual power plants, smart homes, intelligent transportation and smart cities.

Things, in the IoT sense, can refer to a wide variety of devices such as 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, or field operation devices that assist firefighters in search and rescue operations. Legal scholars suggest regarding "things" as an "inextricable mixture of hardware, software, data and service". [5] These devices collect useful data with the help of various existing technologies and then autonomously flow the data between other devices. Figure 1 shows a technology roadmap of IoT indicating how IoT grew over the years and what future awaits it.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

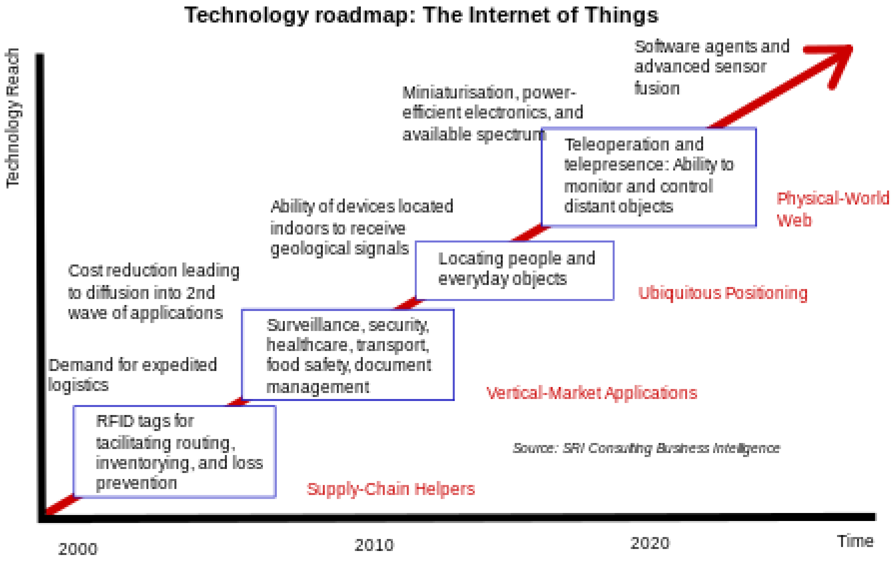


***Figure 1: Technology roadmap for IoT***

The amounts of data generated worldwide every year estimated by different studies [6, 7, 8, 9] are different, it is thought that the total amount of data generated has exceeded one zettabyte in recent years. It is evident that data analysis tools available today are simply not powerful enough to handle and analyze big data of IoT. There is no doubt that it is still a difficult problem to put more than one zetta byte into a single storage system. The next issue we need to take into account is that the data from IoT are generally too big and too hard to be processed by the tools available today. As Baraniuk observed [6], the bottleneck of data processing will be shifted from sensor to the data processing, communication, and storage capability of sensor.

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. [10] It is an essential process where intelligent methods are applied to extract data patterns. Data mining involves six common classes of tasks [10]:

1. Frequent pattern mining – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using frequent pattern mining, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
2. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
3. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
4. Outlier detection – The identification of unusual data records, that might be interesting or data errors that require further investigation.
5. Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
6. Summarization – providing a more compact representation of the data set, including visualization and report generation.

The main focus of this paper is frequent pattern mining and how it can be used for mining data, particularly IoT data.

## II. FREQUENT PATTERN MINING ON IOT DATA

Pattern mining consists of using/developing data mining algorithms to discover interesting, unexpected and useful patterns in databases. Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc. Pattern mining algorithms can be designed to discover various types of patterns: sub-graphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns, high-utility patterns, etc.

The most popular algorithm for pattern mining is Apriori. [11] It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications, including IoT data. The only condition is that the data should be in a transactional data format. A transaction is defined a set of distinct items (symbols). Apriori takes as input (1) a minsup (minimum support) threshold set by the user and (2) a transaction database containing a set of transactions. Apriori outputs all frequent itemsets, i.e. groups of items shared by no less than minsup transactions in the input database.

The Apriori algorithm finds frequent itemsets using an iterative level-wise approach based on candidate generation. The procedure is summarized below:

**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 ∈ D { // scan D for counts
$C_t$ = subset($C_k$, t); // get the subsets of t that are candidates
for each candidate c ∈ $C_t$
c.count++;
$L_k = \{c \in C_k | c.count \geq min\_sup\}$
}
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 $l_2 \in L_{k-1}$
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 $\notin L_{k-1}$ 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 |

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 $\{bread\} \Rightarrow \{milk\}$, 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.

Like most approaches for the association rule, analyzing the purchase behavior still attracts the attention of researchers and companies using the RFID or even IoT approaches. In [12], Schwenke et al. provided a high level method to describe the agent states and the customer behavior in a supermarket, namely, thinking, moving, and action. To solve the problem of customers not being able to find the products they are looking for quickly, in [13], customer specific rules (e.g., age and family state), category-based rules, and association rules are integrated into the same buying suggestion system to help the customers of a supermarket find the products they are looking for. The rule based and case-based reasonings are used to analyze the data to make the system be able to suggest to the customers what they need more accurately based on (1) the personal favor of a customer, (2) the personal purchase history, and (3) the behavior of other people.

The other approach based on the RFID and sensor technologies is the smart environment which also needs mining technologies to make it more intelligent so as to provide more convenient services. Different from most studies on classification that employed the activities of daily living (ADL) to describe the activities, a set of motions are used to facilitate the classification algorithm to differentiate different events. The focus of [14, 15, 16] is on the relations between temporal activities. Inspired by [17], several studies [14] were focused on describing and defining the relations of temporal activities. The relations are first divided into before, after, meets, met-by, overlaps, overlapped-by, starts, started-by, finishes, finished-by, during, contains, and equals, which can then be used to describe the relations between temporal activities.

The apriori algorithm is also used in [14] to discover the frequent patterns. Later studies [18], [19] add the k-means clustering algorithm to classify the activities to create a normal mixture model for each activity before the association rules mining is applied. Another study [20] integrates the association rule mining with the linear support vector machine (LSVM) classification algorithm to improve the accuracy rate of a behavior prediction system for the health care at home. Patterns captured by four different kinds of sensors—the ECG sensor, temperature sensor, network camera for the human location and motion, and facial expression sensor—are used to detect the events. The LSVM is used to recognize the home services while the association rule plays the role of analyzing the home services for the human if any sudden events occurred, such as a human suffering from stress.

## III. CONCLUSION

In this paper, a technology roadmap for IoT was described followed by the description of which data mining techniques can be used on IoT data. Then, it was shown how frequent pattern mining can be done on data from IoT that is in the form of a transactional database. Finally, association rule mining was introduced and it was discussed how it can used on IoT data.

## IV. REFERENCES

[1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks, vol. 10, no. 7, pp. 1497–1516, 2012.

[2] D. Bandyopadhyay and J. Sen, "Internet of things: Applications and challenges in technology and standardization," Wireless Personal Communications, vol. 58, no. 1, pp. 49–69, 2011.

[3] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the internet of things," IEEE Internet Computing, vol. 14, no. 1, pp. 44–51, 2010.

[4] 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.

[5] 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

[6] R. G. Baraniuk, "More is less: Signal processing and the data deluge," Science, vol. 331, no. 6018, pp. 717–719, 2011.

[7] D. Reed, D. Gannon, and J. Larus, "Imagining the future: Thoughts on computing," Computer, vol. 45, no. 1, pp. 25–30, 2012.

[8] M. Hilbert and P. L´opez, "The world's technological capacity to store, communicate, and compute information," Science, vol. 332, no. 6025, pp. 60–65, 2011.

[9] 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.

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

[11] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. ACM SIGMOD International Conference on Management of Data, vol. 22, no. 2, 1993, pp. 207–216.

[12] 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.

[13] S. Somasundaram, P. Khandavilli, and S. Sampalli, "An intelligent RFID system for consumer businesses," in Proc. International Conference on Green Computing and Communications and International Conference on Cyber, Physical and Social Computing, 2010, pp. 539–545.

[14] V. Jakkula, D. Cook, and A. Crandall, "Temporal pattern discovery for anomaly detection in a smart home," in Proc. IET International Conference on Intelligent Environments, 2007, pp. 339–345.

[15] 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.

[16] S. L¨uhr, G. West, and S. Venkatesh, "Recognition of emergent human behaviour in a smart home: A data mining approach," Pervasive and Mobile Computing, vol. 3, no. 2, pp. 95–116, 2007.

[17] J. F. Allen and G. Ferguson, "Actions and events in interval temporal logic," University of Rochester, Tech. Rep., 1994.

[18] 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.

[19] 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.

[20] 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

## CITE AN ARTICLE

Shafi, M., Daiya, S., Dalal, S., & Baja, S. (2018). MINING INTERNET OF THINGS DATA. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *7*(5), 370-374.