Data Mining Framework for IoT Applications

Priyanka Gupta
Assistant Professor
Vivekananda Institute of Professional Studies
GGSIPU, New Delhi

Rajan Gupta, PhD
Associate Professor
Vivekananda Institute of Professional Studies
GGSIPU, New Delhi

ABSTRACT
With the increasing dependency of events over the smart objects which can be easily controlled and monitored, can be identified automatically and communicate with each other through internet and can make decisions by themselves, there is an urgent need of new paradigm which can connect these smart devices together and that new paradigm is Internet of Things. The enormous amount of data produced by IoT devices can be converted into knowledge using data mining techniques. In this paper we analyze a data mining framework for different IoT applications.

Keywords
Internet of Things (IoT), Data Mining

1. INTRODUCTION
Today, every object in the world whether it is sensor, equipment or other devices has become smart, i.e., they can sense the events, make wise decisions and control the environment conditions. So, the concept called Internet of Things is required to automate these objects.

Smart objects are not only the smart phones, tablets or laptops, but they can be anything with a sensor on it like cars, washing machines, embedded devices, security systems, smart watches or even a human being with a heart monitor etc. All these devices use IP address as a unique identifier.

According to S. Haller et al [1] "A world where physical objects are seamlessly integrated into the information network, and where the physical objects can become active participants in business process. Services are available to interact with these 'smart object' over the Internet, query their state and any information associated with them, taking into account security and privacy issues."

The Internet of Things (IoT) refers to the next generation of Internet which will contain trillions of nodes representing various objects from small ubiquitous sensor devices and handhelds to large web servers and supercomputer clusters [2]. IoT integrates the classical networks with the emerging new technologies such as ubiquitous computing, cloud computing, data mining, sensor networks, RFID technology, mobile communication technologies, machine to machine learning etc. From the viewpoint of technology, IoT is an integration of sensor networks, which include RFID, and ubiquitous network. From the viewpoint of economy, it is an open concept, which integrates new related technologies and applications, productions and services, R. & D., industry and market [3].

IoT not only attracts various researchers in different fields, academicians, industry professionals as it is one of the most promising and vast research area but also attracts businessmen at different levels in their organizations as it reduces waste of resources and optimizes the use of technologies to increase the profit of the organizations. IoT generates and captures a large amount of data that is distributed, heterogeneous and very complex. To handle such vast data, big and heterogeneous database systems and data warehouses are required. Also, to analyze such data to identify unknown patterns and to make IoT smarter, data mining technologies are required so that IoT becomes more efficient and useful. Basic data mining algorithms and technologies are not sufficient for IoT framework. So, it becomes a great challenge and responsibility to collect, analyze and manage IoT data and also to generate and update data mining algorithms for IoT purposes.

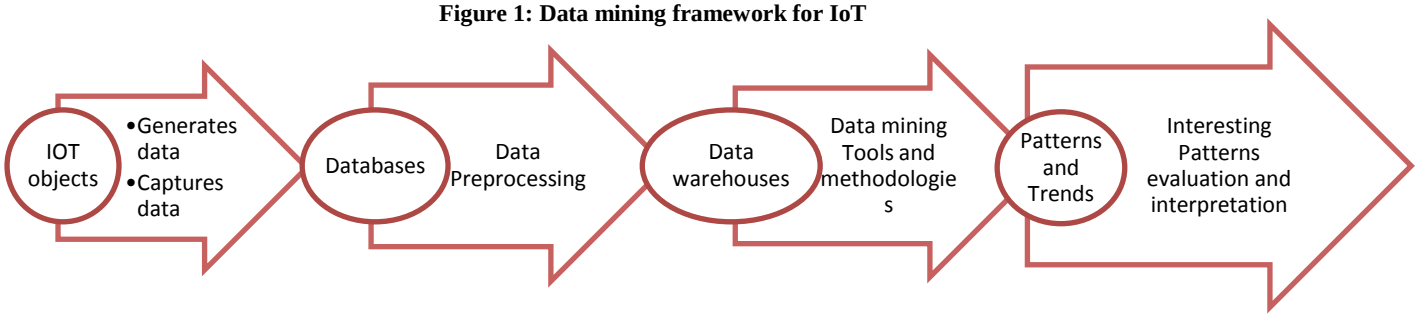
Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [4]. In the era of IoT, where everything interacts and communicates with each other, a large amount of data is produced which must be analyzed and properly mined to enhance the functionalities of IoT. If we can use data mining methodologies in IoT in most effective manner, then this combination proves to be a game changer in the economy of any country.

Data mining techniques should be integrated with IoT for organizational decision making. Hence, this paper is primarily aimed at presenting a detailed description of data mining framework proposed for IoT applications. The framework presented in this paper will provide a path to the researchers interested in solving IOT applications using data mining.

2. DATA MINING FRAMEWORK FOR IoT
Data mining is one of the most powerful and emerging technologies which is used to mine certain useful trends and patterns which are unknown, to enhance the performance of the organizations. Almost all the organizations are growing rapidly with the help of data mining functionalities. Data mining helps in finding out something in the massive data which is unknown and most profitable to the organization. For example, by finding out the frequent buying patterns of the customers, a company can increase sales by placing those items together which are being purchased together, by applying discounts on those items or by reducing redundant items.

The primary aim of Knowledge discovery in databases is to discover the novel patterns in the large data sets. It is the blend of many different domains which includes: artificial intelligence, machine learning and statics. Data mining transforms a data set into understandable structure and extract important information which helps in gaining an insight into the raw data collected from various IoT applications.

Figure 1: Data mining framework for IoT



Therefore, IoT forms a network of physical objects or things which are embedded with electronics, sensor and network connectivity by which the devices can collect and exchange data [5]. The perfect association between IoT and Data Mining results into a new innovative technology which will benefit every section of the society. This association give birth to a number of different applications. These applications generates enormous amount of heterogeneous data. As the data in IoT application is generated continuously from different sources like wireless sensor networks, RFID etc.

Ali et al [6] differentiated types of data from IoT into "data about things" and "data generated by things". Data about things refers to data that describe things themselves (e.g., state, location, identity, and so on) and data generated by things refers to data generated or captured by things. Normally, the former contains data that can be used to optimize the performance of the systems, infrastructures, and things of IoT whereas the latter contains data that are the results of interaction between humans, between human and systems, and between systems that can be used to enhance the services provided by IoT. This new type of data (i.e., data captured by sensor or RFID) has been defined as a kind of "big data" [7]. The huge amount of data generated by IoT applications and potential of KDD motivated us to analyze a data mining framework for IoT applications. Based on the data mining and IOT overview, the data mining in IOT process is as follows: (Figure 1).

Data Mining framework for IoT applications begins with the first step of capturing the tsunami of data generated from IoT objects which includes: Sensor networks, Actuators, WSN (Wireless Sensor Network), WSAN (Wireless Sensor and Actuator Network), RFID (Radio Frequency Identification)Tags, Camera, GPS etc. These objects generates and capture a large amount of data in different forms. For better understanding of different data forms generated by IoT objects, a brief description of three important IoT objects namely Sensor Networks, WSN and RFID is given below.

2.1 Sensor Networks

A sensor network is a group of small devices and equipments which monitor and store the conditions and data in any environment. Today in every industry these networks are used be it retail industry, health industry, manufacturing, scientific and engineering applications, education industry, home appliances etc. it comprises of wireless sensor networks.

2.2 2.2 WSN (Wireless Sensor Networks)

WSN is a network of various spatially distributed autonomous nodes which covers a particular region, gathers data about it and provides the information. There are thousands of wireless sensor nodes which have some computational power, some memory, limited bandwidth and sensing capability. They capture and store the data of the environmental conditions and send the data to a sink or a base station which further processes and analyzes the data. The sensor nodes can communicate with each other using radio signals. Some real time applications of WSN include transportation, military, area monitoring, healthcare etc.

2.3 2.3 RFID

RFID means Radio Frequency Identification. It used to identify objects and people using radio technology of short range. These objects and people also can communicate with each other digitally. There are 2 entities here: Reader which is stationary and Tag which is movable. This tag on any object is being read by the reader to identify it. RFID systems consist of three components in two combinations: a transceiver (transmitter/receiver) and antenna are usually combined as an RFID reader. A transponder (transmitter/responder) and antenna are combined to make an RFID tag [8]. The real power of RFID comes in combination with a backend that stores additional information such as descriptions for products and where and when a certain tag was scanned.RFID readers scan tags, and then forward the information to the backend. The backend in general consists of a database and a well defined application interface. When the backend receives new information, it adds it to the database and if needed performs some computation on related fields. The application retrieves data from the backend [9].

Example of RFID systems and implementation is supermarket chains. All the products is the shopping bag of the customer may be identified by RFID reader as all the products have RFID tags on them. After identifying the products, the information is send to the backend which further provides the additional information such as prices of the products, discounts on the qualifying products and for the qualifying customers, etc. Also the backend reduces the number of products in the database and notifies the concerned authority if the stock of the products have to be updated.

In many applications, barcodes are often used but the difference between bar codes and RFID systems are RFID systems do not need a line of sight, RFID readers may scan many products at one time and they have more storage than barcode systems.

Once the data generation and collection is over, the second step of the framework is Data Preprocessing. Data preprocessing is coined as manipulation and enrichment of data by some authors [10]. The massive data produced by these IoT devices emerges the need for databases to store them. To analyze such large amount of data, data warehouses are required for which data preprocessing is needed which includes cleaning the data (removing noisy, inconsistent and incomplete data), transforming the data which includes

converting the data into the forms appropriate for data analyzing, reducing the data which includes removing the redundant data and building the aggregate data where required.

Third step of the framework is selecting an appropriate data mining methodology for converting the preprocessed data into knowledge. The combination of big data with IoT requires distributed network. Data created by these distributed networks is very difficult to analyze. The current data mining methods are more efficient to run on single network. They need to be modified in order to run through IoT framework. Currently only small scale IoT systems' data can be mined. So, it is a big challenge to transform those methods to a large scale IoT distributed systems.

Minimizing energy dissipation and maximizing network lifetime are among the central concerns when designing applications and protocols for sensor networks. Clustering has been proven to be energy-efficient in sensor networks since data routing and relaying are only operated by cluster heads. Besides, cluster heads can process, filter and aggregate data sent by cluster members, thus reducing network load and alleviating the bandwidth. [11]

2.4 Clustering for IoT

Clustering algorithms [12] divide data into meaningful groups so that patterns in the same group are similar in some sense and patterns in different group are dissimilar in the same sense. Here, we do not have a training data. This is unsupervised learning technique. E.g., search engine uses clustering method to group several web pages into different groups like news, videos, images, blogs etc.

Clustering is an efficient way to enhance the performance of IoT on the integration of identification, sensing, and actuation. That is why many new clustering algorithms are developed for the WSNs that are probably the most common devices to be found on the IoT.[13] One of the most well-known clustering algorithms that take into consideration the energy conservation of WSN is apparently the low-energy adaptive clustering hierarchy (LEACH).[14]

2.5 Classification for IoT

Classification is an important technique in data mining as it involves predicting the output based on the given training data. This is called supervised learning technique. In this technique, a training data is given by the use of which a classifier model is build and based on this model the future pattern is predicted. E.g., a classifier model is used to predict whether a loan application of an applicant must be passed or not based on the previous data of previous customers.

For IoT applications, classification algorithms are divided into 2 types – outdoor and indoor. The example for the outdoor IoT is traffic jam problem. The smart phones and other smart devices are used to predict the traffic situations and communicate them to the users. Some classification algorithms are there to solve the traffic jam problem, where the classifiers predict and suggest the less jammed route to the driver based on the previous traffic conditions.

The example for the indoor IoT is smart home. There are so many sensor technologies, web cameras, microphones and infrared presence sensors applied to build a smart home. They are used to monitor the space for smart homes. To analyze the data produced by these devices, several classification methods are there to detect and predict the activities and to track the human behaviors inside smart home.

The result of data mining methodologies are certain patterns and trends out of which we have to find out the interesting patterns which are profitable to the organization. Hence, the final step of the framework is Interesting Pattern Evaluation. After applying the data mining methodologies, a large number of patterns are evaluated. All of these patterns are not interesting. So, we have to evaluate the interesting patterns which are the most profitable to the organization. Patterns become interesting when it is totally unknown till yet and not expected. E.g., in the consumer products company, if association analysis is being performed, then the pattern of purchasing bread and ball is more interesting than the pattern of purchasing bread and butter in the morning. This is because the first pattern is unexpected and new to the user while the second pattern is the most common. So, to evaluate the interestingness of the patterns there are mainly the 2 types of measures –

2.6 Objective measures

These measures are the first filters to identify interesting patterns. The structure and statistics of patterns play an important role in objective measures. Here, previous knowledge about the data is not required. These measures depend upon the theories in probability, statistics and information theory. Examples of such measures are support, confidence, classification error etc. Support measure reduces the dataset to be examined. Confidence measure helps in selecting the strong rules out of all the association rules.

2.7 Subjective measures

These measures are the second filters to identify interesting patterns. These measures depend upon the user's beliefs and background knowledge about the data. The 2 main subjective measures are as follows:

2.7.1 Unexpectedness

When the pattern doesn't match with the user belief, then it is called unexpected. Sometimes such patterns become interesting because they are surprising to the user. For example, in a product supply chain, suddenly the demand of a particular product decreases to 8% in last 1 year but otherwise its demand was very high before the last 1 year, usually 80-90%. So, this pattern is surprising as well as most interesting to the user since it alters the user's expectations which are based on the user's beliefs.

2.7.2 Actionability

Here, a pattern is considered as interesting when the user can do something about it which means he/she can take further action to use that interesting pattern to increase the profit of the organization. E.g., if a good employee's performance falls below a certain threshold, then he/she may be warned to increase his/her performance.

2.7.3 Novelty

A pattern becomes novel when it is unknown and very new to the user. If a user knows the pattern or can derive the pattern from the existing patterns, then it is not novel. A novel pattern always become interesting because it adds to the knowledge of the user. E.g., a user knows the rule: if (used_seat_belt='yes') then (injury='no'). So, it is not novel. But, if the new rule is discovered which is: if (used_seat_belt='yes') and (passenger='child') then (injury='yes') then this rule is novel because it is new to the user.

3. CONCLUSION & FUTURE SCOPE

IoT generates enormous amount of heterogenous valuable data. To convert this data into knowledge, data mining systems are developed. In this paper, we presents a data mining framework for IoT applications. This paper identifies three challenges for IoT and data mining. Since, IOT produces a large amount of data in terabytes(TB), petabytes(PB) or zettabytes(ZB), to perform data mining on such complex and large data, efficient data preprocessing methods must be applied. Secondly, the knowledge to be mined is not straightforward. It is deeply hidden in the data. So data mining tools must be updated to mine such complex knowledge from IOT objects. Finally, Security of IOT data is a big challenge while performing data mining as sometimes there is much sensitive information in the data like banking transactions, medical records, military data etc. So, strong and strict security mechanisms must be applied while performing data mining on such sensitive data.

4. REFERENCES

[1] S. Haller, S. Karnouskos, and C. Schroth, "The Internet of Things in an enterprise context," Future Internet Systems (FIS), LCNS, vol. 5468. Springer, 2008, pp. 14-8.

[2] Anne James, Joshua Cooper, Keith Jeffery, and Gunter Saake. "Research Directions in Database Architectures for the Internet of Things: A Communication of the First International Workshop on Database Architectures for the internet of things (DAIT 2009)," BNCOD2009: 225-233.

[3] Zhang Lin. "School of Management, Zhejiang University, Prof. Liu Yuan: The business scale of communications between smart objects is tens of times the scale of communications between persons," Science Times. 2009.11.16. (in Chinese)

[4] H. Jiawei and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2011.

[5] Ravi Uttarkar and Raj Kulkarni, "Internet of Things: Architecture and Security", in International Journal of Computer Application, Volume 3, Issue 4, 2014

[6] N. Ali and M. Abu-Elkheir, "Data management for the internet of things: Green directions," in Proc. IEEE Globecom Workshops, 2012,pp. 386–390.

[7] S. Berkovich and D. Liao, "On clusterization of "big data" streams," in Proc. International Conference on Computing for Geospatial Research and Applications, 2012, pp. 3:1–3:1.

[8] Susy d'Hont, "The cutting edge of RFID Technology and Applications for manufacturing and Distribution", Texas Instrument TIRIS, 2002.

[9] Christoph Jechlitschek, "A survey paper on Radio Frequency Identification (RFID) Trends", Reports on Recent Advances in Networking

[10] Chien C.-F. and Chen L.-F.. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with Applications Vol 34, Issue I, Jan 2008 pp. 280-290.

[11] A. Chamam and S. Pierre, "A distributed energy-efficient clustering protocol for wireless sensor networks,"Computers & Electrical Engineering, vol. 36, no. 2, pp. 303 – 312, 2010.

[12] A.K.JainandR.C.Dubes, Algorithms for Clustering Data,Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[13] D. Uckelmann, M. Harrison, and F. Michahelles, "An architectural approach towards the future internet of things," in Architecting the Internet of Things, 2011, pp. 1–24.

[14] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energyefficient communication protocol for wireless microsensor networks," in Proc. Hawaii International Conference on System Sciences, 2000.