International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
National Conference "NCPCI-2016", 19 March 2016
Available online at www.ijrat.org

Data Mining for Various Internets of Things Applications

Krushika Tapedia, Anurag Manohar Wagh

Abstract—Internet of Things is now an accelerating technology in the world of devices. It helps us connect all the devices which we use in our day to day chores via the internet. Starting from home, office, industry automation to health care and smart cities internet of things has revolutionized the world by interconnecting them. As a result it generates massive volumes of data. For many this data has immense business value and information. This is where data mining comes into play which makes such kind of systems smarter enough for better efficiency and greater opportunities and services. This paper introduces to the Internet of Things technology and states the need of data mining in a world where everything is delivered over the internet and explains the process and suitable algorithms required for Internet of things.

Index Terms— Data mining, Internet of things, Knowledge Data Discovery.

I. INTRODUCTION

THE Internet of Things (IoT) is an emerging technology whose basic idea is to connect all the physical devices.
The following is the definition of Internet of things given by S. Haller et al. [1] IoT: "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." In a world where people are trying to develop machines which can think on their own, data mining has proven to make an IoT system smarter, and also a prominent reason behind the success of IoT.

[2] It is also anticipated that by 2020, the amount of internet connected things will reach 50 billion, with $19 trillion in profits and cost savings coming from IoT over the next decade. These smart commodities connected via the internet could be sensor networks, RFID technology, and various handheld or mobile devices.
The data produced by these commodities is huge in volume. It can be justified by considering an IoT system for temperature and humidity monitoring of a garden or a farm. Here the temperature and the humidity detecting sensors are connected all over the place and these sensors send data every hour. Now let us imagine this scenario for a day and that each sensor sends one mega bytes of data per day, and so what if there are 100 such sensors in a farm. The data collected will be in enormous amounts for a system, and humongous for a larger system. To maintain and generate some valuable business information out of it, also provide various services to enhance the development and planning of the system data mining is necessary. Now the challenge is to make this system smarter , what if this system helps the farmers to predict the climate given by the temperature and humidity sensors , an efficient graph is plotted reporting various attributes like soil moisture information , by pH sensors report the acidic level of soil which can help the farmer to decide whether or not to use fertilizers, irrigate only at a particular place by checking if the water level is low which is given by the data procured on the moisture sensors and the graphical report of the level of water in the soil by minimizing water wastage and land clogging. The main intention behind citing this whole scenario was to show how data mining is making this low cost system so efficient and easily manageable. There are various processes and algorithms in data mining, so to select a particular algorithm for a particular IoT system is also a challenge now.

Krushika Tapedia is studying in the Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
(email: krushika29@gmail.com)
Anurag Manohar Wagh is studying in the Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
(email: anuragmwagh@gmail.com)

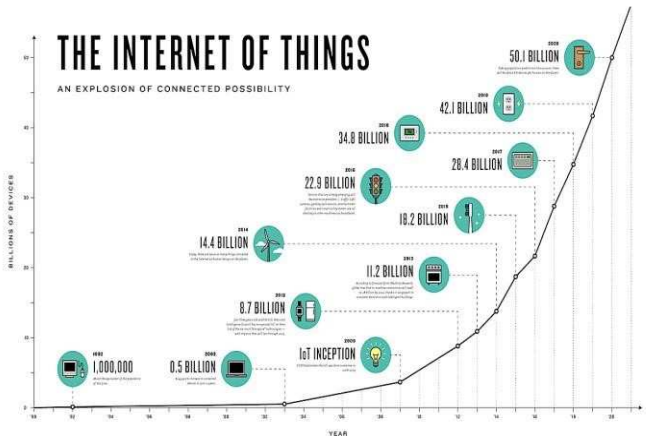

Fig. 1. Here we can see the growth of devices connected over the internet which is increasing rapidly over a couple of years.

On a further note we could also add that for bigger systems such as Super markets using RFID technology. Each product requires 18 bytes of raw data to be stored and there are almost 700,000 products in a super market chain. If the RFID Reader scans the products per second, the data produced will be almost 12.6 GB per second and it will reach about 544 TB per day [4]. This produces tremendously large data over a year which could be in PB (Peta Bytes) that is called as Big Data. There are many challenges and research issues going on in big data and data mining such as the assimilation of heterogeneous data sources and data types like the data gathered from sensors, social media, cameras etc., also the data in various formats such as byte, string, binary and so on.

The fact that huge volume of data is produced in an IoT system is incontrovertible by the above mentioned examples and scenarios. So it is quite evident that there is a need of Data mining for these interconnected systems i.e., IoT.

II. DATA MINING

Data mining in simple terms can be said as the process of extracting valuable or sensible information from a huge set of data using patterns or the relationship between the data to generate revenue or sometimes to cut costs also. Data mining is also exemplified as Knowledge Data Discovery (KDD), many state that KDD and Data Mining are indistinguishable also many consider that data mining to be an indispensible step in KDD.

A simple process model used popularly in data mining will be discussed in this section and also how this is reliable to implement for all the IoT systems with a basic suitable model will also be discussed further.

A. Data Mining Processes

There are two ways in which the processes of the data mining is explained one is the KDD processes with seven stages where as the other process model is the Cross Industry Standard Process for Data Mining (CRISP – DM) which has six stages inclusive of Business Understanding as the name suggests this process model deals with the Industry standards so a basic business understanding is inevitable as traditionally companies are mined to see future trends and better opportunities in the company.

For solving our present scenario which is to manage the huge data from IoT and apply suitable data mining technique, we will first look up the seven stages in the KDD process
which are as follows –

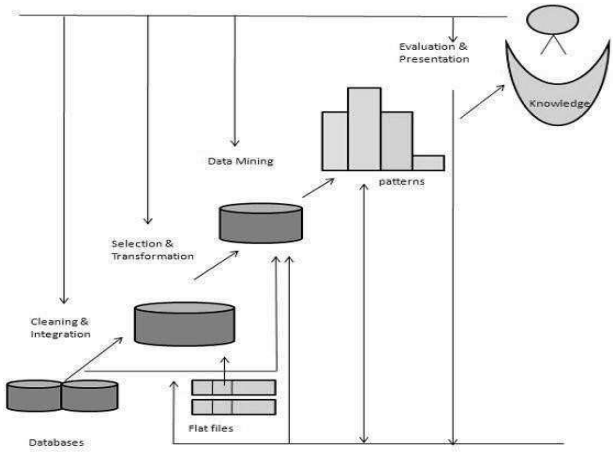

Fig. 2. The figure depicts the basic process model of Knowledge Data Discovery which comprises of cleaning, integration, selection, transformation of data followed by pattern evaluation and presentation.

- Cleaning: The erratic data which has no role in providing valuable information is to be removed.
- Integration: This process is to associate various types of data.
- Selection: In this step the pertinent data is to be restored from the database to achieve proper knowledge by analyzing appropriate data.
- Transformation of data: The term transformation itself states that there is a change in the state of data, i.e., the data's format is changed from the source system to the destination system by performing various operations on it such as mapping or summation.
- Data Mining: As mentioned above, this step is to extract information from the database on the basis of the required patterns using suitable algorithms.
- Evaluation: Through which pattern the data is being extracted and information is generated is evaluated to ensure the correctness of the information.
- Presentation: Finally, the information required is plotted in the form of graphs or other statistical methods for better understanding.

The above mentioned seven stage KDD processes are the typical process stages under which data mining is performed. Further discussion is upon how this model is suitable for IoT.

B.Suitable Data Mining Processes for IoT

We live in a world where the speed with which the business needs to move is much faster than the time it takes to conceive and launch new solutions in the areas of big data, data mining, cloud, and IoT [3]. To find relatively small chunks of data in peta byte sized databases generated from an IoT system is like looking for a black cat in a coal cellar. To get in the game, variety of data mining algorithms should be built with various capabilities to get insights and reduce the risk of project failures. Till today there are many studies which have been trying to solve the problem of acquiring of big data on IoT systems. Most of the mining techniques are developed to execute on a single system, so these KDD systems cannot be applied directly to process big data of the IoT system, whereas for a small system undoubtedly these KDD processes can be applied directly.

To develop a high geared data mining structure of KDD for an IoT system the following three points [5] are to be considered to elect the suitable mining technology, and they are –

- First and the foremost it is essential to understand the definition of the problem, their limitations and required information and so forth.
- Secondly, the major concern would be to understand what kind of data is to be required like the representation, size of data, processing of different data etc.,
- Thirdly on the basis of the above mentioned points, a suitable data mining algorithm is to be chosen to bring out sensible and required information from the raw data.

Further the types of data mining algorithms are being explained.

C. Data Mining Algorithms

- Classification: It is a function of data mining that delegates items into categorical labels. It helps us to predict the category of a particular item in a dataset. Let's consider a scenario where a marketing manager of an automobile company wants to analyze the probability of a customer buying a type of car on the basis of his/her profile. A classification model can be utilized to predict the type of car; family, sports, truck or van, that a customer is likely to buy on the basis of his/her age and family background.

  There are various classification models such as decision tree, neural networks, IF - THEN rules depending upon their use.

- Clustering: Unlike classification, clustering is typically defined as categorizing the data into some sensible, meaningful groups or classes. This helps to achieve an easy perceptive for the users by grouping naturally. The best example for this could be a search engine which is based on clustering, that can categorize endless web pages into news, images, videos, reviews etc.,

  There are various clustering models such as kMeans clustering, k-Medoids clustering, Densitybased clustering and Hierarchical clustering that can be used depending upon their use.

- Association Analysis: Market basket is the best relatable module to association. Market basket analysis is observed routinely in supermarket chains where the items which are likely to be bought together with another set of items are always placed together such as toothbrush and toothpaste are always in the same section. This helps in decision making. At first the data is processed incessantly, for first catalog of association analysis.

  To discover inter transactional association apriori algorithm has been used followed up with association discovery. Other algorithms used are pattern growth, event-oriented, event-based, partition based, FP Growth, Fuzzy set and incremental mining.

- Time Series Analysis: When data points are present in consecutive time interval, time series analysis is applied to extract meaningful related to specific patterns or statistics. Stock market index value is analyzed in a time series manner. Time series analysis is also used in forecasting, to analyze dependent events; that is to predict future values based on past events.
- Outlier Detection: Intermittently there exists a data which is not complaisant with general behavior or model of the data. This kind of data is different from remaining set of data which is called as outlier. This type of data contains useful information regarding aberrant behavior

of the system comprised of outliers. Outlier analysis can be used to extrapolate outliers, to calculate distance among objects, distribution of input space.

The above mentioned data mining functionalities with the listed algorithms are the most commonly used algorithms in any field to mine the data and extract the required information.

III. Relating IoT Applications and Data Mining

As there is a rapid growth in the devices and sensors connected over the internet, we have a treasure trove of applications in this field. Some of the successful applications are listed below.

A. Smart City

The various IoT systems in a smart city are discussed below relating it to the appropriate data mining functionality used to make the system better and smarter.

1) Traffic Control:

IoT devices such as GPS, smart phones, vehicle sensors deployed across the city can provide data points such as travel time, frequency of heavy vehicles, accident prone zones and construction areas. These data points can provide the insights to the reason behind congestion in the targeted area. Here, classification algorithm can be used to solve traffic congestion problem. Targeted areas can be classified depending upon the high, medium, low probability of occurrence of traffic jam in a particular area. Classification model can be used to predict the time of the day where the congestion will be at the peak and alternative route can be used to arrive at the destination. This will distribute the traffic and avoid congestion.

2) Residential E-meters:

Conventional meters are being rapidly replaced with smart meters as smart meters can provide real time data about the energy consumption in digital format through email or even on smart phones. However, Time Series analysis on time series data, which is automatically gathered at different intervals throughout the day can be used to predict energy consumption, provide notifications by any means if any anomaly is detected in energy consumption. Synthetic data can be generated from available real data, which can be used for forecasting.

3) Pipeline Leak Detection:

Maintaining water pipe leaks for municipal corporations is a cumbersome job. Especially with old pipes, with the use of sensors sound of water passing through can be analyzed using outlier detection algorithm to identify leaks. Hence, taxing job of detecting water leaks can be simplified and in addition, cost of maintenance can be reduced to the half as compared to the conventional method.

B. Home Automation

Data generated by IoT devices used in home automation can be mined to generate meaningful patterns. These patterns can be used to predict future events and provide automated interaction with the user. Home automation requires classification and time series analysis models. Where interactive devices are connected together can be classified upon their usage. Data generated by these devices can be stored with their corresponding timestamps, this data can be used in forecasting to predict occurrence of an event at a particular time, using linear regression.

C. Health Care

The improvement in health care industry is evidently seen due to the advancements of IoT systems in it. These IoT systems offer innumerable services for users to check on their health such as medication adherence systems, calorie burnt, blood pressure, blood glucose, heart rate, weight measuring devices and pulse oximeters and store the data on some cloud based platforms maintained by required hospitals. In order to make these intelligent a system should be developed to integrate these heterogeneous data and give accurate information about the patient. The patient doctor specific prescriptions and medical history can be text mined and draw important conclusions about the present condition of the patient[6], chances of survival of the patient, and [7] clustering can be done for the better treatment and care of the patient. We could also outlier it to identify any unusual patterns which will be easy in detection of any fraud.

IV. Conclusion and Future Scope

In this paper, we have discussed about the new emerging technology that is Internet of Things (IoT), later moving on to how data mining is an important part of IoT which makes these systems smarter by discussing the general processes of data mining. Also we have seen key points to keep in mind while selecting an appropriate algorithm for an IoT system. Further discussion was about the widely used data mining functionalities with their specific algorithms and various IoT applications relating it to the suitable data mining functionality applied to enhance the system for better services.

Finally, an IoT system which has the potential to acquire proper insights from these huge oceans of data available are reliable in today's fast pacing world.

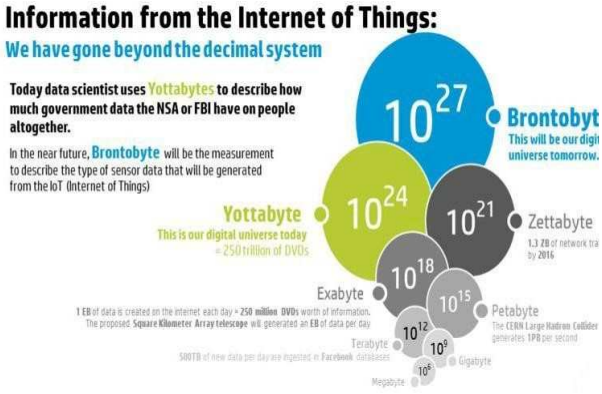

Fig. 3. This figure depicts the future of Internet of Things.

- Big data analytics for IOT software revenues will experience strong growth, reaching $81 billion by 2022 says Strategy Analytics
- Smart Cities will use 1.6 billion connected things in 2016
- By 2025 IOT will be a $1.6 trillion opportunity in Healthcare alone
- 50 billion+ connected devices will exist by 2020 Data captured by IOT connected devices will top 1.6 zetta bytes in 2020

Expansion in devices connected to Internet of Things is growing rapidly and a sheer volume of data is being generated. This data can help predict accidents, crime, provide doctors with real time statistical reports about patient's health, productive maintenance on equipment in industry, pipelines in the cities. However, in future the conundrum will be running out of ways to analyze big data created by IoT devices. Machine learning is a "subfield of computer science (CS) and artificial intelligence (AI) that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions" [8].

Machine learning in IoT can be utilized to take Tera Bytes of data and narrow it down to meaningful data. Better decision s can be made by collecting pattern and similarities that can be learned by machines. Anomalies can be detected using machine learning algorithms. Any object which is part of IoT architecture can be trained effectively to perform more than one specified operation.

Artificial Intelligence is the best way to understand the data generated by IoT devices without AI we fall to the risk of Relevance Paradox. When a company seeks information only relevant to them Relevance paradox occurs. But, there may be information in (its widest sense, data, perspective, general truth etc.,) that is not perceived as relevant because information seeker doesn't already have it and its relevance becomes apparent only when seeker acquires it. Thus, information seeker is trapped in paradox.

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] Plamen Nedeltchev,"It is inevitable. It is here. Are we ready?" *The Internet of Everything is the new Economy* [Online]. Available : http://www.cisco.com/c/en/us/solutions/collateral/enterprise/ciscoon-cisco/Cisco_IT_Trends_IoE_Is_the_New_Economy.html

[4] Shen Bin, Liu Yuan, Wang Xiaoyi, "Introduction to Research on Data Mining", *Research on Data Mining Models for the Internet of Things* [Online].Available: https://www.ceid.upatras.gr/webpages/faculty/vasilis/Courses/SpatialTemporalDM/Papers/InternetOfThings05476146.pdf

[5] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang, "Basic Idea of Using Data Mining for IoT,"*Data Mining for Internet of Things: A Survey*, IEEE Communications Surveys & Tutorials, vol.16.

[6] L. Duan, W. N. Street, and E.Xu, "Health Care Information Systems: Data Mining Methods in the Creation of a Clinical Recommender System," *Enterprise Information Systems*, vol.5, no.2, pp.169-181, 2011.

[7] B. K. Schuerenberg, "An information excavation. Las Vegas payer uses data mining software to improve HEDIS reporting and provider profiling," *Health Data Management*, vol. 11, no. 6, pp. 80–82, 2003.

[8] Theo Priestley, *Series of Unfortunate tech Predictions – Artificial Intelligence & IoT are inseparable* [Online]. Available: http://www.forbes.com/sites/theopriestley/2015/12/08/a-series-of-unfortunatetech-predictions-artificial-intelligence-and-iot-areinseparable/2/#2bce1cb67bdd