Data Streaming of Healthcare from Internet of Things (IoTs) using Big Data Analytics

Dost Muhammad Khan Assistant Professor, Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Punjab, Pakistan. Email: khan.dostkhan@gmail.com

Muhammad Jameel Sumra Visiting Lecturer, Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Punjab, Pakistan.

Faisal Shahzad Lecturer, Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Punjab, Pakistan.

Abstract

The present study aims at the concept of the IoTs (IoT) and its relation with the healthcare sector. Nowadays, IoT is the main focus of researchers and scientists while this concept illustrates the data stream generated from IoT devices in massive amounts like big data with a continuous stream that requires its proper handling. This study aims at the analytical processing of big datasets having a medical history of patients and their diseases. The data cleansing is applied before going through the analytics phase due to the existence of some noisy and missing data. The analytics of data identified that what events are happening while the mining approaches identified why and how events are happening. Together, both phases help in data analytics and mining. Finally, the analytics and visualization led to the decision making and its results depict the effectiveness and efficiency of the proposed framework for data analytics in IoT.

Key Words

IoTs (IoT); Big Data; Health Care; Data Analytics

Introduction

A healthcare IoTs gadget makes predictable surges of information, and affiliations must be able to handle the high volume of stream data and perform analysis on that data. IoTs is the idea which has been under thought for a long time and analysts are investigating the capacities, restrictions and further headways in it. All through some late years, there is a pattern of investigating IoTs idea around the world. The real idea of IoT delineates the world where computing machines would facilitate the general population in different aspects of data and other information with the guide of some robotized ways. The idea of the web of things is to make the world where everything will be associated with the web and as in result, a web is formed surrounding us in such a way that we are intercommunicated with each and every object. The principle center is around the IoTs gadgets from the human services associations which are producing the information floods of patients (Ilayaraja, 2013). The information is, for the most part, put away in the printed version shape yet in current innovative patterns, the human services associations are likewise changing as per the necessities of the present time (Ahokangas, Juntunen and Myllykoski, 2014). The possibility of the web of things is to make the world where each object around us, from vehicles to home machines and diverse devices, will be connected to the web and would make requests for a server and as in result, request a response (Akash et al., 2014).

The data is generally secured in the printed variant shape yet in current creative examples, the human administration's affiliations are moreover changing according to the necessities of the present time. The conceptual IoTs framework on the premise of cloud computing is visualized in Fig 1:
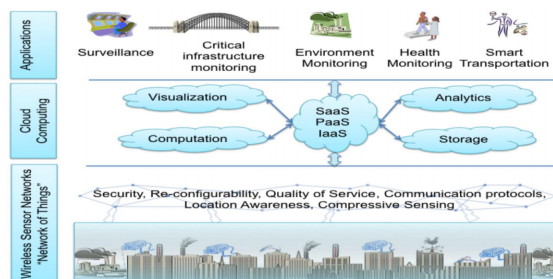
Figure 1. Conceptual IoTs framework with Cloud Computing at the center (Gubbi et al., 2013)

While selecting the headway stack for colossal information dealing with, the immense joining of information that the IoTs will pass on must be remembered. Affiliations ought to adjust degrees of progress to unequivocally misuse IoTs information and control of all systems and methodology will be impacted. There are various edges that should be seen as when IoTs with the setting of human administrations are discussed; one of the rule things that surface at the highest point of the need rundown is the possibility of a gigantic, consistent stream of information hitting affiliations' information stockpiling. Holding this under thought, server stockpiles must be set up to handle this kind of circumstance when an extra heap of heterogeneous information would be secured (Chen et al., 2012).

A human services IoTs device makes unfaltering surges of data, and affiliations must have the ability to handle the high volume of stream information and perform practices on that information. On the premise of necessities and issues discussed more than, an answer is being proposed to look at the data of patients from the human administration's affiliations IoTs devices. The importance of IoTs in the last few years is demonstrated in Figure 2.
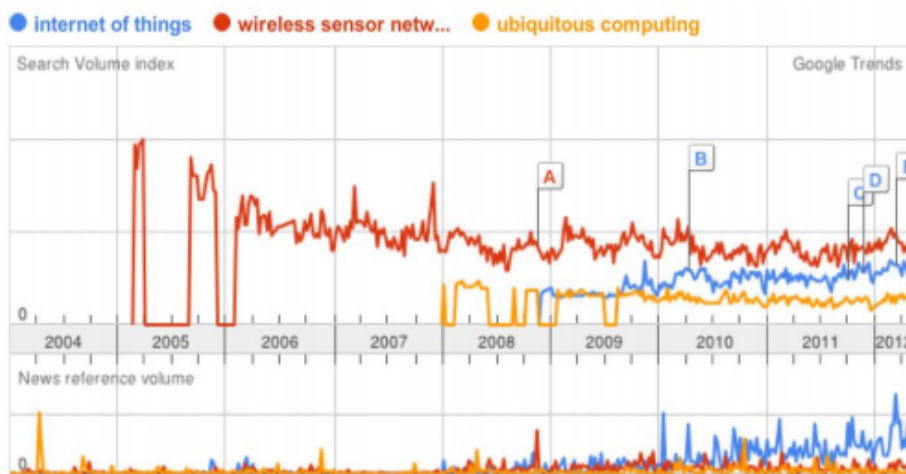
Figure 2: Search trends for Google in the last few years (Gubbi et al., 2013)

The healthcare and medically-based systems are working in the remote nations while they have a rough size in petabytes. On the basis of necessities and issues talked about over, an answer is being proposed to examine the information of patients from the human services associations IoTs gadgets. The information would be broke down to make the capacity effective by mining the data which is pertinent and expelling the garbage information from that stream.

While selecting the advancement stack for gigantic data taking care of, the huge joining of data that the IoTs will pass on must be recollected. Affiliations conform headways to explicitly exploit IoTs data and control of all frameworks and procedures will be influenced. There are numerous angles which ought to be viewed as, for example, when IoTs with the setting of human services is talked about, one of the principal things that surface at the top of the priority list is the idea of a huge, constant stream of data hitting associations' data stockpiling. The rest of the paper is organized as Literature is reviewed in Section 2, in Section 3 is about Methodology, the Results are discussed in Section 4 and finally, the conclusion is drawn in Section 5.

Literature Review

Many researchers and scientists have proposed various techniques in the field of big data analytics. Data analytics can be performed in various ways. There are diverse edges which should be seen as, when IoTs with the setting of human associations is discussed, one of the deciding things that surface and no more amazing motivation behind the need once-over is the probability of a huge, steady stream of information hitting affiliations' information stockpiling (Khan et al., 2013). The data may vary in formats and can require preprocessing as well. While selecting the advance stack for huge information dealing with, the creature joining of information that the IoTs will pass on must be recalled.

Rajaraman, 2016 reported in his study that representation of data helps human picking and keeping fundamental information yet this can be an invigorated approach. Information ousted from data would be further dismembered computationally using cushioned measures systems and self-organizing structures to end up with exceptional decisions as to information. The mining of essential information or representations is performed near to the examination and a specific evaluation will be done to diagram the mined and separated qualities and information.

The data is required to be envisioned so that the information bound can be further helpful in decisions taking. The impression of data is truly a basic and principal time of data examination to address the multifaceted information expelled from data to different people. This accomplice is envisioning data, quantifiably, using unmistakable instruments.

Raghupathi and Raghupathi, 2014, discussed the relationship between healthcare and big data. An answer had been proposed on the premises of building a structure to depict the capacities of huge information investigation in the field of social insurance frameworks. Crude information from an enormous information source is changed to organized information utilizing information warehousing approaches and facilitate investigation is performed on the premises of huge information examination instruments. Later on, the information had been mined utilizing a few information mining approaches. The review introduced an underlying adaptation of an expository framework to address the speed volume test of the enormous information stream from an IoTs gadget, and results and discoveries had been talked about keeping the brilliant city idea, producing abnormal state necessities for a major information logical framework. The stage has been named as the mist registering stage and the interruptions of IoTs in regards to dangerous multiplication are examined alongside a progressively appropriated design stage. The fog registering stage fundamentally concentrates on IoTs, its enormous information examination and the expansion from the edge of the system.

Akash et al, 2014, talked about a revolutionary innovation in terms of a smart wheelchair as an IoTs device. The rule spin is around the IoTs gadgets from the healthcare which are conveying the information surges of patients. While selecting the progress stack for enormous data managing, the monster joining of data that the IoTs will pass on must be recollected. Affiliations should conform degrees of advance to unequivocally abuse IoTs data and control of all frameworks and procedures will be influenced. This also created a link of IoTs with healthcare. The paper provided a thorough review of advance and wearable computing, connected with the IoTs service and generating streams of data. The ecosystems had been discussed on the premises of their core functionality such as data creation, information extraction, the pragmatics of information and actions on the base of this pragmatics.

Pang et al., 2015 discussed the IoTs and in-home healthcare services. The data from remedial organizations can help in clinical decision really strong systems that can go about as specialists while encountering the essential data of patients with similar signs and ailment. The data from restorative organizations can help in clinical decision really strong systems that can go about as specialists while encountering the fundamental data of patients with relative signs and disorder. While selecting the advance stack for huge information dealing with, the beast joining of information that the IoTs will pass on must be remembered. There are diverse edges that should be seen as, when IoTs with the setting of human associations are discussed, one of the deciding things that surface and no more bewildering reason for the need once-over is the probability of a colossal, consistent stream of information hitting affiliations' information stockpiling. Holding this under thought, server stockpiles must be set up to handle this kind of circumstance when an extra heap of heterogeneous information should be secured.

Pang et al., 2013 discussed the IoTs with healthcare and ecosystems. A human services IoTs device makes continuously streams of data, and affiliations must have the capacity to handle the high volume of stream information and perform the investigation on that information. The human organizations and remedially based structures are working in remote countries while they have an unforgiving size in petabytes. On the introduction of necessities and issues discussed more than, an answer is being proposed to look at the data of patients from the big data IoTs devices. The data would be penniless down to make the limit equipped for mining the information which is associated and removing the decline data from that stream. The examination kept under thought is cloud-based. This would help in investigating and sifting important data from enormous information streams and would wind up in insightful and strong choice takings for brilliant urban communities set. This also involves human administration. The main objective of the research is to refine data from a stream of healthcare to IoTs devices through analytics and to make big data storage of IoTs devices from healthcare effective and efficient

Methodology

The proposed methodology focuses on healthcare organization devices. The devices focused in this study are wearable devices for remote monitoring and which can track the patients' activity and diagnose disease through different operations and predictions. The IoTs device generates a continuous stream of data which makes this a stream of Big data. The proposed methodology emphasizes on the analytics of such data and is comprised of various phases.

The basic methodology for analytics using big data stream initiates from the collection of data which may be actually the source of data, IoTs devices. The data is subjected to some preprocessing, analytics and post-processing to aid in decision making and extraction of information from all the IoTs data. The pipeline of the methodology is displayed in Figure 3 below.
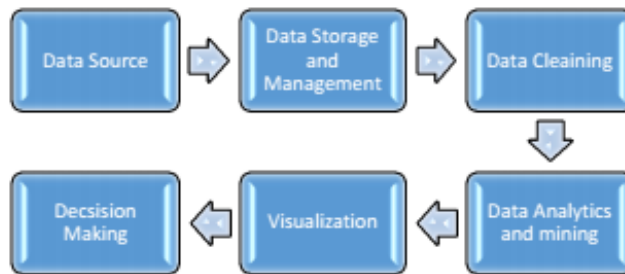
Figure 3: The pipeline of Data Analytics Methodology

Data at each phase is processed using different tools of IoTs big data streams and integrations of all premises of data integration services. The tool used for the whole analytics of the data is Talend Open Studio. Talend offers an eclipse based framework for data management. Pentaho Solutions are used for data integration at various phases and for data cleansing, OpenRefine is used. Data mining is accomplished by Rapid Miner which has numerous data mining techniques. Data Analytics is done through Talend in which data is analyzed in a multidimensional way. The data is acquired from the HealthData repository (healthdata.gov). The data source of this data set is actually a wearable IoTs device that kept on tracking the patients' medical status. The whole medical diagnosis on the basis of that tracking is saved. The medical diagnosis of complete history and records are maintained in the dataset. The data is a classified data set which aided in the further supervised learning. The data is stored and analyzed in a specific format using the Talend. Hadoop is technology adopted for the sake of big data storage, offers specific formation of data and through this approach, data from the IoTs device is stored in the Hadoop-specific format.

Data come up in any shape and before mining and analyze data, it is necessary to clean the junk from data. The data cleansing assured that the data would be consistent and uniform so that the analytics can be carried out smoothly. This process is taken as pre-processing of analysis and mining. The discretization and normalization are also performed on the data. Missing values are handled using the concept of replacement with default values while the outliers are also handled using the normalization of the data within the desired range. The preprocessing of the data also led towards the instance reduction and the dataset size is reduced without affecting the quality of the data. Noise reduction and un-uniform behavior are also handled using the OpenRefine.

The data is analyzed at this stage using the Talend Open Studio Data Analytics Module. The data is subjected towards a multidimensional analysis which helped in analyzing the data according to different aspects and the pivoting of the data is also assured. The symptoms helped in performing the future medicinal predictions and led towards the diagnosis of the disease. The medicinal history of the previous disease is also observed and data is analyzed on this perspective that what precautions and treatments have been taken for specific diseases. Through the whole perspective, it is observed that what kind of events have been occurring in the data and for this scenario, the events highlighted are the medical diagnostics. The most necessary dimension in this scenario is the time dimension which helped in the diagnostics of disease on various timestamps.

The data mining is performed at this phase using the RapidMiner. The data mining is considered as the post-process of the analytics. The analytics does not make sure that why those symptoms are occurring. RapidMiner provided different modules for the mining of data using the Nearest Neighbor (NN), Classification and Clustering, Genetic Algorithms, Machine learning approaches, and the Association rules. The serial association rule mining is performed on the basis of the Apriori technique. The prior performed with quite robustness and effectiveness while it is observed that the technique produced too many candidate itemsets. It is required to perform too many passes over the whole data. The technique used for the item set counting is the dynamic Item set counting technique. Moreover, the FP-Growth technique reduced the candidate generation in the item set finding. As FP-Growth helped in the reduction of the item set, still it is observed that the FP-Tree construction is quite a time taking and hectic. The evaluation criteria based on the chi-square illustrates the strong association rule mining and the rules present in the dataset.

After data analytics, the extracted information is visualized which is helpful in decision making. Data visualizations are an important phase to represent complex information extracted from data which helps in visualizing data, statistically, using different tools. Data is visualized according to different aspects, having a multidimensional view in the form of graphs and other representations to further aid in decision makings. The patterns, associated with the disease are highlighted and visualized which helped in the clear representation of the rules. The visualization of data helps humans deciding and keeping relevant information but this can be a hectic

approach. Keeping this aspect in view, information extracted from data is further analyzed computationally using fuzzy rules systems and self-training systems to end up with final decisions regarding information. The decisions are made using the decision tree and this approach associated each disease with a drug and some treatments.

Results and Discussion

For experimentation, data is initially divided into subparts. Against each chunk, the performance evaluation is performed. The details are provided below in Table 1.

Table 1: Data Load Performance

Data Chunk	No of Nodes	Raw Data Load Time (in minutes)
2GB	2	3.2
	4	2.8
	6	2.9
	8	3.1
	10	2.82
4GB	2	3.6
	4	3.6
	6	3.81
	8	3.92
	10	3.73

Against the data chunks of 2GB and 4GB, it is observed that the loading time has a relation with the data size, while its association with the number of nodes for data is also analyzed. It can be observed that the execution time goes down by increasing the number of nodes while with the increase in the data size, the time of execution also increases. The results are shown in Figure 4.
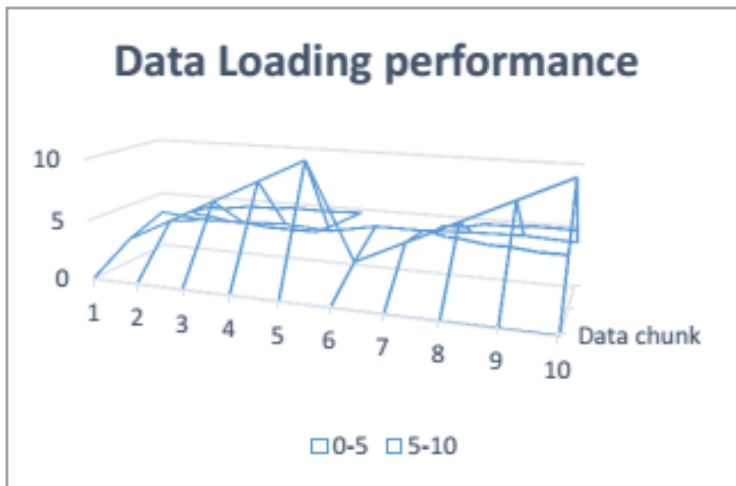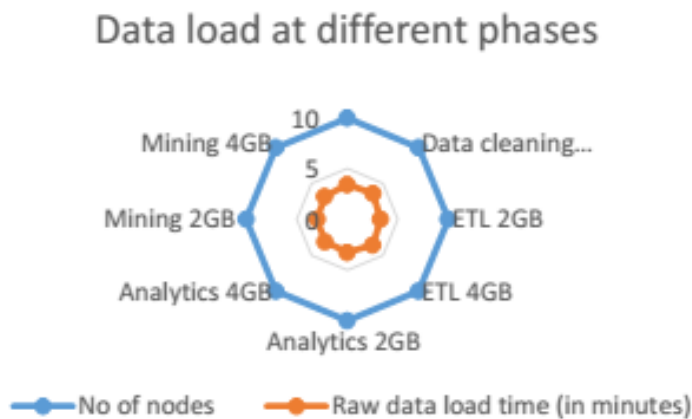

Figure 4: Execution Performance Plot

This plot illustrates the relation between the number of nodes, the data size and its execution time. Among different variants of nodes and data size, it is observed that the data should be kept at a maximum number of nodes as dividing the data into subgroups makes it specifically easy and efficient to be handled. At different phases of the whole experimentation, the data is also observed to be loaded in a different number of minutes. The number of nodes being considered for all the further steps is 10 while both data parts, i.e. 2GB and 4GB are used in all the phases of data experiments. The details of data loading at different phases of the analytics are shown in Table 2.

Table 2: Data Load Performance at Different Phases

Experimentation Phase	Data Chunk	No of Nodes	Raw Data Load Time(in minutes)
Data cleaning	2GB	10	3.42
	4GB	10	3.61
ETL	2GB	10	3.31
	4GB	10	3.59
Analytics	2GB	10	3.27
	4GB	10	3.11
Mining	2GB	10	2.98
	4GB	10	3.16

To clearly elaborate on this comparison, the radar graph technique is used to plot the graph for data load performance execution. The radar graph illustrates the comparison framework and is used for the multidimensional analysis itself. It provides comparison under multiple dimensions as can be seen in Figure 5.

Figure 5: Execution Performance Radar Plot

Through this analytical approach, it is observed how the data size and number of the node are also affecting each phase in IoTs data analytics. Diverse behavior in data is observed while going through all the data phases. In data cleaning, various patterns of data noise are observed. Different data columns are observed to have diverse data in nature. Cleaning is performed on whole data to handle the missing and noisy values. The missing values are observed to be present in the patients' records mostly in age and disease-related information. The age is calculated as in the difference of date of birth and current date. Diagnostic related blanks are left until the data mining stage. The frequent pattern mining at the data mining phase is performed using two approaches i.e. serial mining and parallel mining and it is observed that parallel association mining is more effective and efficient as compared to serial mining. The multidimensional analysis is performed on the premises of disease, personnel, medicines and the time dimension. The conceptual elaboration of the whole multidimensional model is shown in Figure 6. The data analytics is performed on the basis of all the dimensions and different data trends are analyzed so that different events can be highlighted. On the basis of events, the mining is performed to diagnose the reasons for the data events. Through this analysis, we get to know what is happening while data mining elaborates on the main reasons for those events.
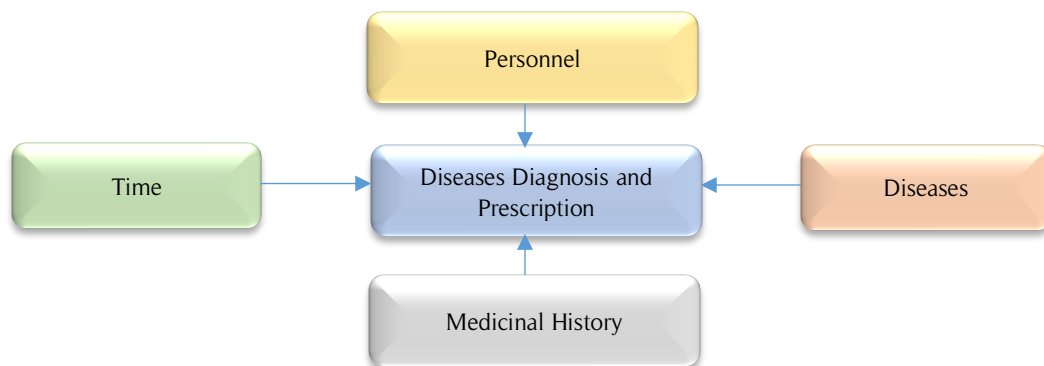
Figure 6. Multidimensional Model

Frequent patterns mining is performed using parallel association rules mining. For visualization of rules, symptoms are associated with the diseases in 5 sets i.e. a, b, c, d, and e. Through this association rule mining, predictions are generated and decisions for future symptoms are made. The association led to the development of a list containing the diseases and the symptoms associated with those diseases. With the aid of the Chi-Square evaluation method, the association with the highest support is separated and the results are shown in Table 3.

Table 3. Disease Association and Evaluation

Problem	Support	Confidence	Chi square	Interest	Conviction
AIDS	20	16.13%	4228.06	212.75	1.19
B12 deficiency	186	18.62%	5815.46	32.48	1.22
Cardiac catheterization	65	42.76%	6597.73	102.79	1.74
Cardiac transplant	72	47.37%	15974.05	222.76	1.9
Congestive heart failure	351	11.51%	3290.76	10.66	1.12
COPD	223	41.22%	5430.53	25.68	1.67
Crohns disease	54	45.38%	4373.43	82.35	1.82
Cystic fibrosis	13	76.47%	12206.84	939.93	4.25
End stage renal disease	39	22.16%	3742.9	97.43	1.28
Glaucoma	218	57.07%	5044.5	24.4	2.27
Gout	495	45.50%	5513.09	24.17	1.8
HIV positive	73	90.12%	10525.03	145.06	10.06
HIV/AIDS	108	87.10%	13584.49	126.62	7.7
Interstitial cystitis	13	44.83%	4617.13	356.52	1.81
Multiple sclerosis	124	34.25%	4166.33	35.02	1.51
Myasthenia gravis	10	25.64%	3451.86	346.68	1.34
Parkinson's	47	33.57%	4377.79	94.56	1.5
Prolactinoma	20	24.10%	4710.94	236.94	1.32
Psoriasis	130	38.92%	3303.25	26.85	1.61
Rheumatoid arthritis	219	38.76%	4405.36	21.5	1.6
Schizophrenia	31	45.59%	5088.16	165.47	1.83

Sickle cell anemia	13	44.83%	8502.43	655.23	1.81
Stress test	63	41.45%	10390.04	166.04	1.7
Systemic lupus	204	23.26%	5863.47	30.02	1.29
Tineacapitis	5	21.74%	4191.79	839.78	1.28
von Willebrand's	7	63.64%	3876.9	555.09	2.75

The evaluation illustrates the diseases which have been associated with various symptoms with the highest support. This illustrates that these diseases can be diagnosed through the proposed framework with maximum accuracy. The graphical plotting of these results is shown in Figure 7. The human organizations and remedially based structures are working in remote countries while they have an unforgiving size of data. On the introduction of necessities and issues discussed more than, the answer being proposed is to look at the data of patients from the big data IoTs devices.
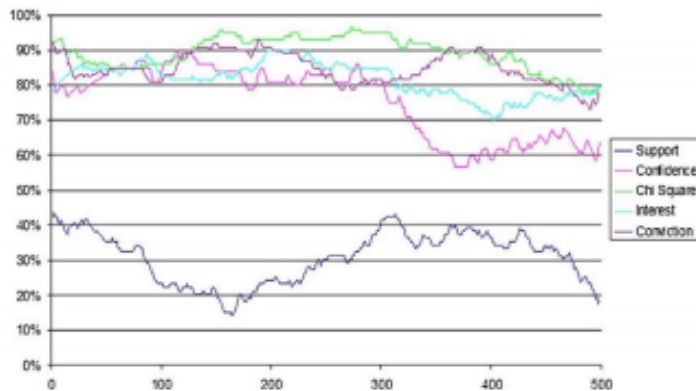

Figure 7: Graphical Illustration of Association Rule Mining Results

The evaluation criteria based on the chi-square illustrates the strong association rule mining and the rules present in the dataset. The results clearly illustrate the effectiveness and efficiency of the proposed framework for data analytics in IoTs. The problem being targeted is health diagnostics and prediction. Various diseases and their symptoms are present in the dataset. The efficiency of data loading has also been computed to specifically analyze the data load of the whole framework. The healthcare sector is a growing sector in the current era.

Conclusion

This study is based on the IoTs (IoT) applications in the healthcare sector. It has been observed that healthcare is one of the strongly growing industries in the world and people are robustly reliable on the growth of healthcare. The present study aimed at the analytical approach in the IoTs data comprising of the patient medical history and further helps in the effective diagnosis of the diseases. The results show the layout the suitability and viability of the proposed structure for the data examination in IoT. The issue is centered around is prosperity diagnostics and desire. Diverse disease and their indications are accessible in the dataset. The efficiency of data stacking has moreover been figured to unequivocally separate the data load of the whole structure. The therapeutic administration's region is a creating division in the present time frame.

References

Ahokangas, P., Juntunen, M., and Myllykoski, J. (2014). Cloud computing and the transformation of international e-business models. A Focused Issue on Building New Competences in Dynamic Environments, *Research in Competence-Based Management, Volume-7 Emerald Group Publishing Limited*, 7, 3-28.

Akash, S.A., Menon, A., Gupta, A., Wakeel, M.W., Praveen, M.N. & Meena, P. (2014) September. A novel strategy for controlling the movement of a smart wheelchair using IoTs. *In Global Humanitarian Technology Conference-South Asia Satellite (GHTC-SAS), IEEE*, 154-158

Alptekin, N. Y. (2013). The Effect of Clustering in the Apriori Data Mining Algorithm: A Case Study. *Proceedings of the World Congress on Engineering*, 3.

Fernandez, F., and Pallis, G. C. (2014). Opportunities and challenges of the IoTs for healthcare: Systems engineering perspective. *In Wireless Mobile Communication and Healthcare (Mobihealth), EAI 4th International Conference, IEEE*. 263-266.

Gubbi, J, R., Buyya, S., Marusic & Marimuthu. P. (2013). IoTs (IoT): A vision, Architectural elements, and future directions . *Future Generation Computer Systems, Elsevier*, 29(7), 1645-1660.

Hu, F., Xie, D. & Shen, S., (2013). On the application of the IoTs in the field of medical and health care. *IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. 2053-2058.

Hu, R. (2010). Medical Data Mining Based on Association Rules . *Computer and Information Science*

Jia, X., Wang, J., & He, Q. (2011). IoTs business models and extended technical requirements. *In IET International Conference on Communication Technology and Application (ICCTA 2011)*. 622-625.

Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. *IEEE Computer Graphics and Applications, 33(4)*, 20-21.

Khan Z, Anjum A, Kiani SL. (2013) Cloud based big data analytics for smart future cities. *In Proceedings of the 2013 IEEE/ACM 6th international conference on utility and cloud computing*, 381-386.

Pang, Z., Zheng, L., Tian, J., Kao-Walter, S., Dubrova, E., & Chen, Q. (2015). Design of a terminal solution for integration of in-home health care devices and services towards the Internet-of-Things. *Enterprise Information Systems, 9(1)*, 86-116.

Raghupathi, W. & Raghupathi, V., (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1).

Teece, D. (2010). Business models, business strategy and innovation. *Long Range Planning, 43 (2-3)*: 172 – 194.

Uckelmann D, Harrison M, & Michahelles, F. (2011). An architectural approach towards the future IoTs. *In architecting the IoTs*, 1-2.

Voros, J. (2012). Macro-prospection: Thinking about the future using macro-and Big History. *Proceedings of Global Future*.

Xia, F., Yang, L.T., Wang, L. & Vinel, A., (2012). IoTs. *International Journal of Communication Systems*, 25(9). 1101.

Zott, C., Amit, R., & Massa, L. (2011). The business model: recent developments and future research. *Journal of management, 37(4)*, 1019-1042.