JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Context Aware Knowledge Discovery Using Hadoop Cluster in Cloud Environment

Dr.E. Uma
Assistant Professor(Sl Gr)
Department of Information Science and technology
College of Enginering, Anna University, Chennai,Tamilnadu, India.
E-Mail: umaramesh@auist.net

Mahendran E
Research Scholar
Department of Information Science and technology
College of Enginering, Anna University, Chennai,Tamilnadu, India.
E-Mail: mahendran.e@gmail.com

Abstract

Context Awareness is promising technologies that provides health care services and enrich area of big data paradigm. The drift in Knowledge Discovery from Data refers to a set of activities designed to refine and extract new knowledge from complex datasets. The proposed model facilitates a parallel mining of frequent item sets for Ambient Assisted Living (AAL) System of big data that reside inside a cloud environment. We extend a knowledge discovery framework for processing and classifying the abnormal conditions of patients having fluctuations in Blood Pressure (BP) and Heart Rate (HR) by employing Frequent Item Ultra-metric Tree (FIUT). The evaluation can be shown to deliver a much better estimate in detecting proper anomalous situations for different types of patients. This culminates in augmented accuracy. The very efficacy of Context Awareness lies in its technical feasibility to be applied in day-to-day activities of health sector benefitting the common man. The FIU Tree algorithm has been shown to be a non-failing technique under any boundary-value-conditioned dataset. This project encompasses lots of research and application scope and therefore fit to be carried out.

Keywords

Cloud Computing, Virtualization, Hadoop and Health Care.

Introduction

Traditional healthcare system gained importance by the proliferation of Information Technology (IT). E-Health applications have become the mandate with adherence of medical standards for interoperability to share the medical data. It becomes the onus of the fraternity of IT to work for the progress of Healthcare to improve the patient care, reduce the cost of the treatment and better manageability in the benefit of healthcare providers. Healthcare application has turn out to be one of the most famous applications over the Internet.

Ambient Assisted Living (AAL) system consists of heterogeneous huge amounts of patient specific unstructured raw data every day.

Cloud Computing

Cloud generation shows a fast boom in modern-day marketplace and gives an efficient and flexible way to access sources over internet. Predominant benefit of cloud computing includes on demand services which makes more attractive towards major industries and organizations. Users on cloud computing are generally charged based totally at the usage time for their services. Subsequently the utilization of computing gives a potential point of preference of lessening immense capital and operational expenses. This cloud generation consists of the advantage of virtualization of providing high scalability, availability and reliability. There had been numerous definitions for cloud round the arena and list some of the fundamental definitions.

Cloud software/middleware should constitute modern software development technologies and be available as open source. The cloud middleware should explore advanced capabilities such as monitoring, billing, metering, high availability, user management and dashboard. Cloud essentially gives three sorts of services to be precise Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). Cloud refers to each the applications delivered as services over the net and also the hardware and software within the knowledge centers that give those services. Cloud computing offers different service models as a base for successful end user applications.

Virtualization

Virtualization is a promising technology which provides software environments in the form of virtual machines dynamically. Virtualization is a methodology for dividing the computing resource into multiple execution environments. Virtualization is a method for concealing the physical attributes of processing resources to improve the route in which different frameworks, applications, or end users associate with those resources. Virtualization refers to technologies designed to provide a layer of abstraction among computer hardware structures and the software program running on them. by using presenting a logical view of computing resources, as opposed to a bodily view, virtualization solutions make it viable to do run several operating systems to in parallel on a single critical Processing Unit (CPU). This parallelism has a tendency to lessen overhead costs and differs from multitasking, which involves walking numerous OS.

Big Data

The term 'Big Data' alludes to portrays extensive volumes of high velocity, complex and variable information that require propelled systems and innovations to empower the catch, storage, appropriation, administration, and investigation of the data. The property of big data means increasing number of semi-structured and unstructured data, inclusive of text, pix, videos, graphs, audios and many others. The volume property suggests that the entire amount of data grows unexpectedly. Big Data is becoming even more important than ever in various areas. It’s getting even bigger in health care.

Hadoop

Our put forth work defines that hadoop framework for processing the dataset in distributed system. Hadoop is a Linux platform to run or process a huge volume of data and provide better enhancement towards the performance metrics. The outcome of the proposed work will be identifying legitimate atypical circumstances for various sorts of patients, so the exactness and productivity will be improved.

Related Work

Context Aware System

Abdur Forkan et al (2014) researched in ambient assisted residing strives to ease the day by day lives of human beings with disabilities or chronic scientific situations. Designed a cloud-based model for a context-aware system for ambient assisted living, to perform key monitoring and data-aggregation tasks, necessitating data transmission and computation at central locations. The point of interest here is at the improvement of a scalable and context-aware framework and easing the waft among information series and processing. This system analyzed the issues related to scalability, cost, and support of heterogeneous services using a single model.

Abdur Forkan et al (2015) discussed about context aware monitoring for personalized health-care services in the area of big data application. Knowledge discovery-based technique that permits the context-aware system to evolve its behavior in runtime by analysing large amounts of information generated in ambient assisted living systems and away in cloud vault. The proposed model helps evaluation of huge information inside a cloud environment. It first mines the trends and patterns in the data of an individual patient with associated probabilities and utilizes that knowledge to learn proper abnormal conditions. The results of this learning strategy are then applied in context-aware decision-making approach for the patient.

The dataset from Physionet (2015) MIMIC - II (Multi parameter Intelligent Monitoring in Intensive Care) database is also used, because this contains a large number of samples of multiple vital signs (including SBP, DBP and HR). Some of the MIMIC-II records contain more than 24 hours Intensive Care Unit patient data. From these referred works the distributions of SBP, DBP and HR in a day for different patient categories during different activities are measured

Mulvenna (2011) Ambient assisted dwelling services that provide support for people to remain in their homes are increasingly being used in health-care systems around the arena. Generally, those ambient assisted living services provide additional information though location-awareness, presence- cognizance, and context attention capabilities, such as sensors and actuators in the home of the individual receiving care. In addition there is a need to provide abstract information, in context. There are numerous distinctive viewing alternatives converged networks and the ensuing explosion in information and facts has ended in a new hassle, as these new ambient assisted living services battle to convey meaningful information to different groups of give up users. The article discusses visualization of facts from the attitude of the needs of the differing end user groups, and discusses how algorithms are required to contextualize and convey records across vicinity and time. So one can illustrate the issues, current work on nighttime AAL services for people with dementia is defined.

Sridevi et al (2010) An important feature of remote monitoring applications is to identify the abnormal conditions of a patient accurately and so send appropriate alerts to the care givers. In traditional systems, situations are classified by generalized medical rules.

Weider et al (2015) research work is to build an application system for early identification of diseases The application framework is an extremely accommodating device for the health care service providers to improve both quality and productivity in human services. . The application framework is worked around Naïve Bayes (NB) classification algorithm to analyze deteriorating health conditions, readmission rates, treatment optimization and adverse events. They are aiming to provide a model to medicinal services and patient.

Noriss (2006) the analysis of the BP and HR data distribution, a normal distribution is identified when data is limited to a specific activity (e.g. BP is higher when the patient is awake and lower when he/she is asleep). The utilization of typical conveyance in engineered information era is additionally solid for validation because the generated data are nearly similar to real data. The advantage of using such distribution for biomedical data analysis is also proven in some previous works.

George Suciu et al(2015) citizens may suffer from a number of diseases, including but not limited to mild dementia and cognitive disabilities, which require either the institutionalization or the constant support from care-givers. as a rule an observing framework is required that can accumulate and prepare information from different sensors about the healthcare condition of the user and environment parameters. They depict a cloud-based methodology for observing the healthcare condition of citizens and the fusion of big data from heterogeneous records flows coming from the sensors. Furthermore, because context understanding is not easily done with a single source of metadata, They examine metadata available from diverse online resources, aiming to recognize the context of its customers, in an effort to offer them personalized eHealth offerings.

Venkatesh et al (2012) is to build up a protected and wellbeing basic Ambient Assisted Living (AAL) environment which can screen the patient's circumstance and give timely updates. So as to satisfy all these requirements, a keen situation has been made to efficaciously and insightfully control patients' needs. The middleware standard favored for the development and deployment a bevy of ambient and articulate services is Open Service Gateway Initiative (OSGi).

Doan B. Hoang et al (2010) Mobile Cloud for Assistive Healthcare (MoCAsH) as an infrastructure for assistive healthcare. Besides inheriting the benefits of Cloud computing, MoCAsH embraces important concepts of mobile sensing, active sensor records, and collaborative planning by deploying intelligent mobile agents, context-aware middleware, and collaborative protocol for efficient resource sharing and planning. MoCAsH addresses security and privacy through deploying selective and federated P2P Cloud to protect records, maintain facts possession and strengthen factors of protection. It also addresses various quality-of-service issues concerning critical responses and energy consumption.

Cloud System

Amazon Cloud (2013) is a collection of remote computing services, also called web services that together make up a cloud computing platform by Amazon.com since 2006. The most focal and surely understood of these administrations are Amazon EC2 and Amazon S3. The service is advertised as providing a large computing capacity (conceivably numerous servers) much speedier and less expensive than building a physical server ranch.

Suciu et al (2015) defined a cloud-based method for tracking the healthcare circumstance of senior residents and the fusion of huge records from various data flows coming from sensors. Frequently, however a solitary source was not adequate to identify a circumstance and, much more, to expound on it. The The hassle become resolved by taking data from different sources (social networks, feeds, etc.), semantically enriching, aggregating and fusing it to provide better stage context data.

Data Mining System

Tsay et al (2009) an efficient method; the frequent items ultrametric trees (FIUT for mining common itemsets in a database. FIUT uses a special frequent items ultrametric tree (FIU-tree) structure to upgrade its effectiveness in acquiring incessant itemsets. All common itemsets are generated by checking the leaves of each FIU-tree, without traversing the tree recursively, which drastically reduces computing time. FIUT was compared with FP-growth, a well-known and widely used algorithm, and the simulation results showed that the FIUT outperforms the FP-growth.

Yaling Xun et al (2015) present parallel mining algorithms for frequent itemsets lack a mechanism that allows computerized parallelization, load balancing, information distribution, and fault tolerance on large clusters. As a strategy to this hassle, they have designed a parallel frequent itemsets mining algorithm known as FiDoop the use of the Map Reduce programming version. To obtain compressed storage and keep away from conditional pattern bases, FiDoop includes the frequent items ultrametric tree, instead of conventional FP trees.

FiDoop has distinctive features. In FiDoop, the mappers independently and concurrently decompose itemsets; the reducers perform combination operations by constructing small ultrametric trees as well as mining.

A. K. Dey (2000) efficient processing of this large volume of medical, ambient and media data using computational power of cloud infrastructure, extraction of right context information. Data Mining is the efficient discovery of valuable, non obvious information from a large collection of data.

Bo Wu et al (2008) association rule mining is to discover affiliation connections among substantial information sets. Mining frequent patterns is an important aspect in association rule mining. In this paper, an efficient algorithm named Apriori-Growth based on Apriori algorithm and the FP-tree structure is presented to mine frequent patterns. The upside of the Apriori-Growth algorithm is that it doesn't need to generate conditional pattern bases and sub-conditional pattern tree recursively.

Xindong Wu et al (2014) A HACE theorem that characterizes the functions of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data transformation.

Sankar K. Pal et al (1992) the multilayer perceptron, using the back-propagation algorithm, and capable of fuzzy classification of patterns is depicted. The input vector consists of membership values to phonetic properties while the output vector is defined in terms of fuzzy class membership values. This permits efficient modeling of fuzzy or uncertain patterns with appropriate weights being assigned to the backpropagated errors depending upon the membership values at the corresponding outputs. Amid preparing, the learning rate is slowly diminished in discrete strides until the network converges to a minimum error solution. The effectiveness of the algorithm is established on a speech recognition problem.

Architecture

The put forth architecture for Context Management System is depicted in Figure 1. The model facilitates a parallel mining of frequent item sets for Ambient Assisted Living System of big data inside the cloud environment.

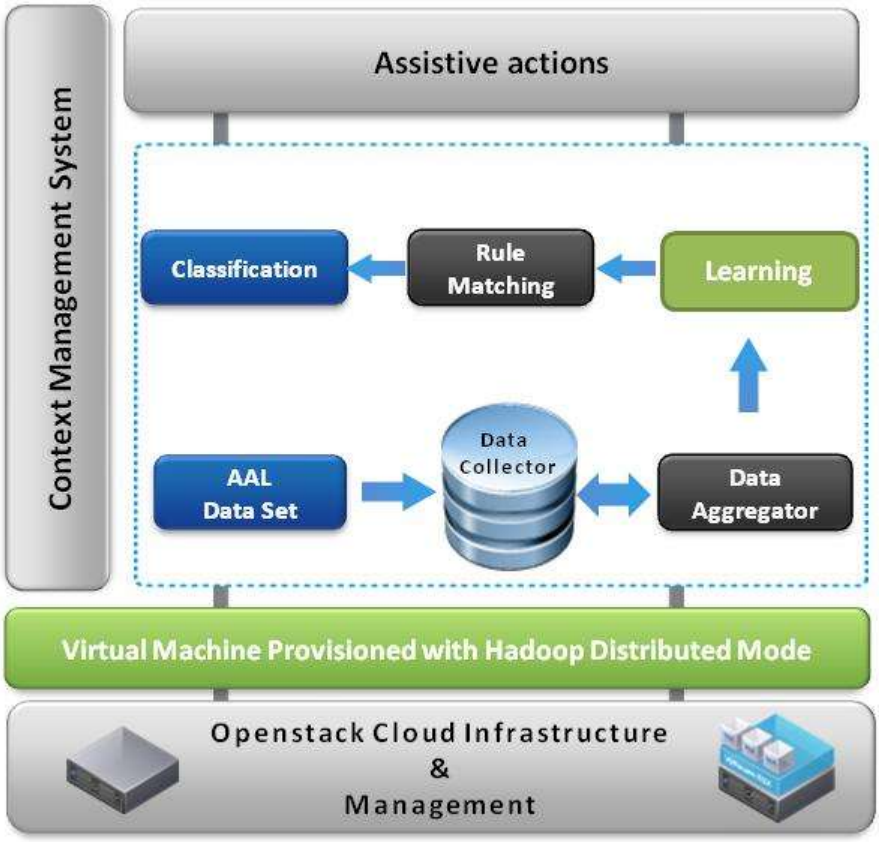

Figure 1 Architecture of Context Aware System with HPC

The knowledge discovery framework for processing and classification the unusual conditions of patients having variations in Blood Pressure and Heart Rate uses Frequent Item Ultra metric tree and storing it in the cloud environment to access anywhere and anytime. The structural design is deployed over the cloud site for management. The architecture primarily consists of Cloud Infrastructure and Management, Virtual Machine with Hadoop and Context Management System.

Architectural Components

Cloud -Infrastructure and Management

The cloud infrastructure has installed with Openstack cloud software Kilo and KVM.

Virtual Machine with Hadoop

A virtual machine is created in openstack cloud to run Hadoop. The provided a virtual machine image containing a preconfigured Hadoop environment. The virtual machine image will run inside of a "sandbox" environment in which we can run another operating system.

Context Management System

A Context Management System (CMS) interacts with different distributed components and binds the operations of each individual component together in the cloud. Context Management is a dynamic computer process that uses 'subjects' of information in one application, to indicate to data occupant in a separate application also containing the similar subject.

Data Set

Data set from Pysionet MIMIC – II contains large no of samples of HR and BP.

Data Collector

Data collector is the process of gathering and measuring information from the data set and storing it in the Database for further actions.

Data Aggregator

Data aggregation is any process in which information is gathered and expressed based on context; In common aggregation purpose is to get more information about particular User based on specific variables such as name, age, health, HR, BP, oxygen level, Etc..,

Learning

The main purpose of the learning system is to perform frequent item set ultra metric tree algorithm of Ambient Assisted Health care data. The FIUT technique employs an proficient ultrametric tree structure, known as the FIU-tree, to display frequent items for mining frequent itemsets. The proposed FIUT comprises of two main phases within two scans of database D. Phase 1 starts by computing the support for all items occurring in the transaction records. Then, a pruning technique is developed to remove all uncommon items, leaving only frequent items to generate the k -itemsets, where the number of frequent items of a transaction is k in a database. Meanwhile, all the frequent 1-itemsets are generated. Phase 2 is the recurring constructions of small ultrametric trees, the actual mining of these trees, and their release.

Rule Induction

Rule generation is a standout amongst the most imperative methods of machine learning. Since regularities covered up in information are frequently expressed in terms of rules, rule generation is one of the basic tools of data mining at the same time.

Classification

Data classification is the procedure of organizing information into category for its best and effective use. A Multilayer Perceptron (MLP) is a feedforward artificial neural network model that maps sets of input information onto a set of suitable outputs. An MLP comprises of different layers of nodes in a directed graph, every layer is fully connected to the next one. Aside from the input nodes, every node is a neuron (or processing element) with a nonlinear activation function. MLP uses a supervised learning procedure called backpropagation for training the network

Assistive Action

Different patient categories considered in the experiment and their average values of vital signs in the analysis of the BP and HR data distribution, a normal distribution is identified when data is limited to a specific activity. The assistive actions are None, Stroke, Syncope and Myocardial Infarction.

Implementation

Openstack Overview

OpenStack is cloud computing software that controls substantial pools of compute, storage, and networking resources all through a data center, all managed through a dashboard that gives administrators control while engaging their users to provision resources through a web interface,

Openstack Compute

Openstack is used for provision and manage large networks of virtual machines. The OpenStack cloud computing environment empowers undertakings enterprises and service providers to offer on-demand computing resources, by provisioning and vast systems of virtual machines. Compute resources are accessible via APIs for developers building cloud applications and via web interfaces for administrators and users. The computer architecture is designed to scale horizontally on standard hardware, enabling the cloud economics companies have come to expect.

Data Set Collection

Collecting Data set from Pysionet MIMIC – II contains large no of samples of height, weight, BSA, BMI, AAL sample dataset is shown in Table 1.

Table 1 Sample AAL Dataset

Record	Gender	Age	Weight	Height	BSA	BMI	Smoker	SBP	DBP	EF	Vascular event
2012	M	72	83	169	1.97	29.06	No	130	75	69	None
2032	F	66	85	160	1.94	33.20	No	150	65	53	None
2033	M	77	82	169	1.96	28.71	No	115	80	42	Myocardial infarction
2119	F	74	91	162	2.02	34.67	No	140	80	72	Stroke

BSA: body surface area (m^2)
BMI: body mass index (kg/m^2)
SBP: systolic blood pressure (mmHg)
DBP: dyastolic blood pressure (mmHg)
EF: ejection fraction (%)
Weight (Kg)
Height (cm)

Data Collector

Data collector is the process of gathering and measuring information from the data set and storing it in the Database for further actions. The schema has been created in postgresql is depicted in Table 2.

Table 2 Postgres Table

Column	Type
Record	Integer
Gender	Character(6)
Age	Integer
Weight	Integer
Height	Integer
BSA	Double Precision
BMI	Double Precision
Smoker	Character(3)
SBP	Integer
DBP	Integer

EF	Integer
Event	Character(10)

In postgres the schema has been created with required fields. The Data set from Pysionet MIMIC – II are inserted in DB.

Data Aggregator

Data aggregation is any process in which information is gathered and expressed based on context; In common aggregation purpose is to get more information about particular User based on specific variables such as name, age, health, BP, oxygen level, Etc..,

The Physionet MIMIC-II data from Database are taken as input and Classifying the information gathered and expressed based on context. To evaluate the performance of the models, the time spent for building each classifier model is measured from small dataset. The classification results obtained for 3 patients are presented in Figure 2

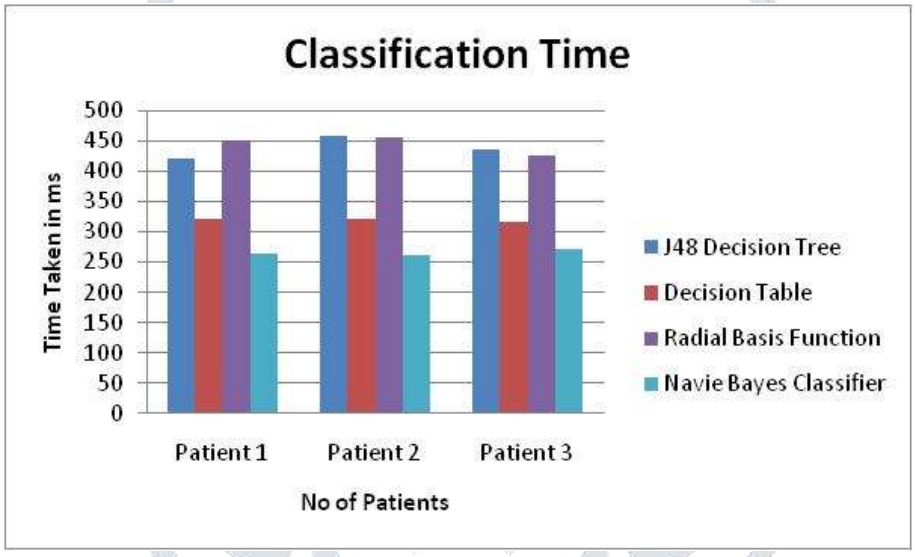

Figure 2 Classification Time

Learning

The most important purpose of the learning system is to perform frequent item set ultra metric tree algorithm of Ambient Assisted Health care data. The FIUT technique employs an efficient ultrametric tree structure, The FUIT searches and traverses recursively.

The processed Physionet MIMIC-II data from Database has been taken for mining of frequent item set. To estimate the efficiency of the FIUT method, several experiments were conducted to compare with the well known FP-growth algorithm. Mining of frequent itemset for 10 patients are depicted in Figure 3.

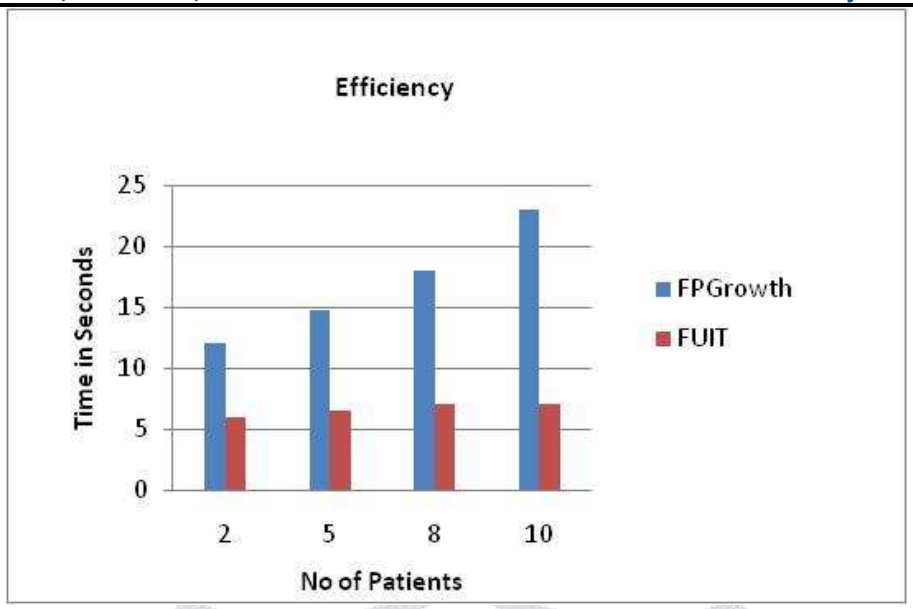

Figure 3 Mining of Frequent Itemset from Database

The time for building the classifiers using MLP for Anomaly detection accuracy and false positive rate for different is depicted in Table 3

Table 3 Anomaly Detection Accuracy and False Positive

Patient	Classification Accuracy		False Positive Rate	
	BDCaM	CAHPC	BDCaM	CAHPC
P1	91.56	92.15	0.117	0.097
P2	95.54	96.78	0.25	0.197
P3	90.75	91.23	0.95	0.893

Conclusion

As of now, a generalized framework for personalized healthcare, this leverages the advantages of context-aware computing, monitoring, cloud computing and big data. The Data set from Pysionet MIMIC-II contains large no of samples of HR and BP are collected, using data collector the data set are stored in the Database for further actions. This provides a systematic approach to support the fast-growing communities of people with chronic illness who live alone and require assisted care. The proposed Context Management System is deployed in Cloud Environment can be given as SaaS with IaaS.

Future Work

The assessment can be appeared as a superior appraisal of recognizing legitimate bizarre circumstances for various sorts of patients. Determining the assistive activity in the project can make it effective with concern to both cost and hazard minimization

Reference

1. A.K. Dey (2000), 'Providing Architectural Support for Building Context-Aware Applications', Ph.D. dissertation, Georgia Institute of Technology.
2. Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari, (2015), 'BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare',IEEE Transactions On Cloud Computing – Accepted for Publication.
3. Abdur Rahim Mohammad Forka, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari,(2014), 'Cocamaal: A Cloud-Oriented Context Aware Middleware in Ambient Assisted Living,' Future Generation Computer Systems, Vol. 35, pp. 114–127.
4. Amazon web services https://aws.amazon.com
5. Bo Wu, Defu Zhang, Qihua Lan, Jiemin Zheng(2008)," An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure", Third IEEE International Conference on Convergence and Hybrid Information Technology, pp 1091-1102.
6. Doan. B. Hoang and L. Chen (2010), "Mobile Cloud for Assistive Healthcare (MoCAsH)," Services Computing Conference (APSCC), IEEE Asia-Pacific, Hangzhou, pp. 325-332.
7. George Suciu, Alexandru Vulpe, Razvan Craciunescu and Cristina Butca, Victor Suciu (2015), 'Big Data Fusion for eHealth and Ambient Assisted Living Cloud Applications', IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp.102-106.
8. Mulvenna M, Carswell, W, McCullagh, P, Augusto, J.C., Huiru Zhen, Jeffers, P., Haiying Wang and Martin, S(2011)., 'Visualization of Data for Ambient Assisted Living Services', in Communications Magazine, IEEE , Vol.49, No.1, pp.110-117.
9. P. R. Norris (2006), "Toward new vital signs: Tools and Mmethods for Physiologic Data Capture, Analysis, and Decision Support in Critical Care," Ph.D. Dissertation, Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA.
10. Physionet MIMIC-II http://www.physionet.org/mimic2/.
11. R - language and environment https://www.r-project.org/about.html
12. S. Sridevi, B. Sayantani, K. P. Amutha, C. M. Mohan, and R. Pitchiah (2010) "Context Aware Health Monitoring System," in Medical Biometrics. Springer, 2010, pp. 249–257.
13. Sankar K. Pal and Sushmita Mitra (1992),"Multilayer Perceptron, Fuzzy Sets, and Classification", IEEE Transactions On Neural Networks, Vol. 3, No. 5, pp 683-697.
14. Suciu G, Vulpe, A., Craciunescu, R., Butca, C and Suciu, V (2015) "Big data fusion for eHealth and Ambient Assisted Living Cloud Applications," IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom) pp.102-106, 18-21.

15. Venkatesh V.; Vaithyanathan, V.; Kumar, M.P.; Raj, P.(2012), 'A Secure Ambient Assisted Living (AAL) Environment: An implementation view,' International Conference on Computer Communication and Informatics (ICCCI), pp.10-12.
16. Weider Pratiksha C, Swati, S, Akhil, S and Sarath, M (2015) "A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment," in Computer Software and Applications Conference (COMPSAC), Vol.1,pp.75-86.
17. Weka 3 - Data Mining with Open Source Machine Learning www.cs.waikato.ac.nz/ml/weka/
18. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding (2014), "Data mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp. 97–107.
19. Yaling Xun, Jifu Zhang, and Xiao Qin (2015),"FiDoop: Parallel Mining of "Frequent Itemsets Using MapReduce", IEEE Transactions On Systems, Man, And Cybernetics: Systems – Accepted for Publication.
20. Yuh-Jiuan Tsay, Tain-Jung Hsu and Jing-Rung Yu (2009),' FIUT: A New Method for Mining Frequent Itemsets', Journal of Information Science, Vol. 179, No.11, pp 1724–1737.