Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 2, February 2016, pg.16 – 23

Feasibility Support of Datamining Tools for IoT

Madhuri.A[1], Deepika.N[2]

[1]Department of CSE, New Horizon College of Engineering, VTU, India

[2]Department of CSE, New Horizon College of Engineering, VTU, India
[1] madhu.ksmile@gmail.com; [2] deepikajvijay@gmail.com

Abstract— Today rapid development in Information Technology field and new applications being developed constantly to support critical businesses all over the world and also various key fields like environment studies, Medicine, Research and other sciences, making it an indispensable part of our lives. Computations across and over Internet has become order of day. Internet of things is an emerging topic of technical, social and economic significance. With IPV6 addressing in place, every possible consumer product, durable goods, cars and trucks, Industrial and utility components, sensors and other everyday object can become part of Internet. Data Analytics is the driving force running these applications and for mining useful knowledge. This document aims at bringing out the feasibility support that current data mining tools render IOT, against its possible data models.

Keywords— Internet of Things (IOT), Data mining (DM), Data mining tools, IoT Data Models, KNIME, RAPIDMINER

I. INTRODUCTION

The term Internet of Things generally refers to scenarios where network connectivity and computing capability extends to objects, sensors and everyday items not normally considered computers, allowing these devices to generate, exchange and consume data with minimal human intervention. Projections for the impact of IoT on the Internet and economy are impressive, with some anticipating as many as 100 billion connected IoT devices and a global economic impact of more than $11 trillion by 2025. There is, however, no single, universal definition. The concept of combining computers, sensors, and networks to monitor and control devices has existed for decades. The recent confluence of several technology market trends, however, is bringing the Internet of Things closer to widespread reality. These include Ubiquitous Connectivity, Widespread Adoption of IP-based Networking, Computing Economics, Miniaturization, Advances in Data Analytics, and the Rise of Cloud Computing.

IoT implementations use different technical communications models, each with its own characteristics. Four common communications models described by the Internet Architecture Board include: Device-to-Device, Device-to-Cloud, Device-to-Gateway, and Back-End Data-Sharing. These models highlight the flexibility in the ways that IoT devices can connect and provide value to the user. Distributed data mining models can solve problems from depositing data at different sites. Data mining model from multitechnology integration perspective describes the corresponding framework for the future Internet.

A. *What is Data Mining?*

Data mining (sometimes called Data or knowledge Discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

II. DATA MINING TOOLS

A, Brief Over view of data mining tools

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use and complexity involved in building data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own advantages and disadvantages. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviours, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. The top six open source tools available for data mining are briefed as below Table 1[2].

TABLE 1
DATA MINING TOOLS

TOOL	**FUNCTION**
Weka	Weka is a Java based open source tool data mining tool which is a collection of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering, and association rule extraction
KEEL	KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. It has a collection of libraries for pre-processing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning, and providing scientific and research methods.
R	Revolution is a free software programming language and software environment for statistical computing and graphics.
KNIME	Konstanz Information Miner, is an open source data analytics, reporting and integration platform. It has been used in pharmaceutical research, but is also used in other areas like CRM customer data analysis, business intelligence and financial data analysis.

RAPIDMINER	RAPIDMINER is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.
ORANGE	Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting.

III. IOT DATAMING MODELS AND FEASIBILITY STUDY

Key features of the proposed models discussed below and feasibility support for IoT models with above mentioned data mining tools.

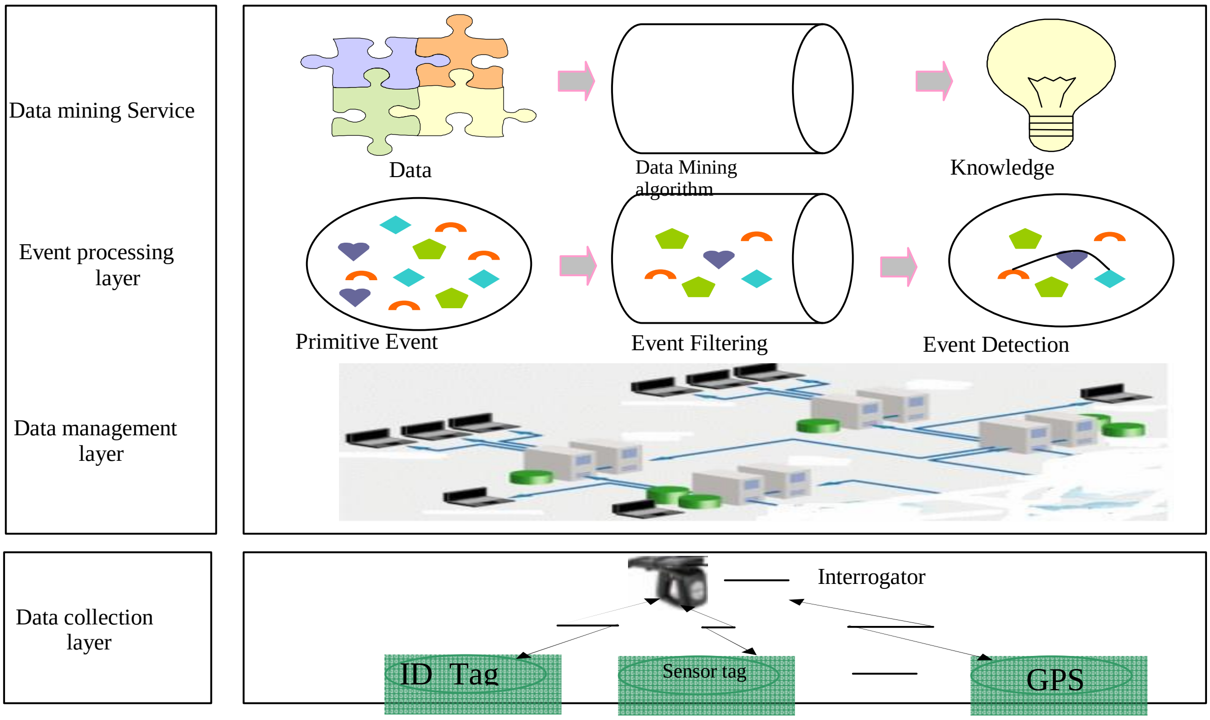

Fig 1 Multi-layer Data mining model

A. Multi-layer data mining model for IoT

Based on architecture of IoT and data mining framework of RFID, this model consists of four layers: data collection layer, data management layer, event processing layer and data mining service layer as shown below in Fig 1[1].Data collection layer adopts devices, e.g. RFID Reader and sinks etc., to collect various smart object's data, Different type of data requires different data collection strategy. Data management layer applies centralized or distributed database or data warehouse to manage collected data. Event processing layer is used to analyse events in IoT effectively.

Data mining service layer allows various object-based or event-based data mining services, such as classification, forecasting, clustering, outlier detection, association analysis or patterns mining, are provided for applications.

TABLE 2
MULTI-LAYER FEASIBILITY SUPPORT

TOOL	**Pros by model**	**Cons by model**
Weka	Weka handles CSV, ARFF, C4.5, binary data files. Can interoperate varied data formats, with multiple data sources. It is best suited for mining association rules, hence can adapt to identify implicit dependency or data in flow from particular sensors more appropriately. With machine learning capability, it can adapt to dynamic changes in configuration. This tool scales well for multi-layer IoT model on the conceptual level, like event processing.	It suffers from scalability issues, to apply to future datasets. Not so polished and CSV reader not as robust as Rapid miner. It suffers from "Kitchen sink syndrome", overhead of additional loads.
KEEL	Limited but provides machine learning tool, to assess evolutionary algorithms for Data mining problems. It contains a big collection of classical knowledge extraction algorithms, Computational Intelligence based algorithms, genetic fuzzy systems and evolutionary neural networks. This tool scales well for multi-layer IOT through blend of support for varied algorithms.	Dynamic adaptability for new algorithm support is restrictive. Hidden or unseen dependency are difficult to handle, if prior not identified.
R	With its support for statistical computing and graphics, it is ideal for handling particular spatially related data. Being a open source tool, it provides an additional flexibility to modify the package configuration on demand. Numerical programming is better integrated in R.	Less specialised for Data mining, but provides extensive statistical library support. Might not scale well with varied applications in IoT model.
KNIME	Custom nodes and types can be implemented in KNIME within hours. Easy to try out and also requires no installation besides downloading and archiving. It can integrate with weka data mining environment and also R-scripts to machine learning, data mining, text mining, predictive analytics and business analytics. It handles complex nested operator chains for huge number of learning problems. It can take the advantage of multi-layer IoT model, as it fits in well most of databases and also includes run, and provide vast statistical routines. It can scale well for data collection, data management and event processing layers of multi-layer model.	No additional support for machine learning. May fail to handle event processing scenarios of trivial and event filtering capabilities of multi-layer data model.
RAPIDMINER	Rapid Miner uses a client/server model with the server offered as a service or cloud infrastructure. It provides an integrated environment for predictive analysis and statistical computing.	More database centric approach, flexibility with other systems could be challenging. Data collection from varied sources might need extra tuning.
ORANGE	It's a component based machine learning suite. It is written in python, hence easier to learn. It has a cross platform GUI. Shortest script for doing training, cross validation, algorithms comparison and prediction. It can fit into multi-layer model for IoT.	Limited list of machine learning algorithms. Reporting capabilities are limited to exporting visual representations in data models. May suffer from space and time efficiency as IoT system scales.

B. Distributed data mining model for IoT

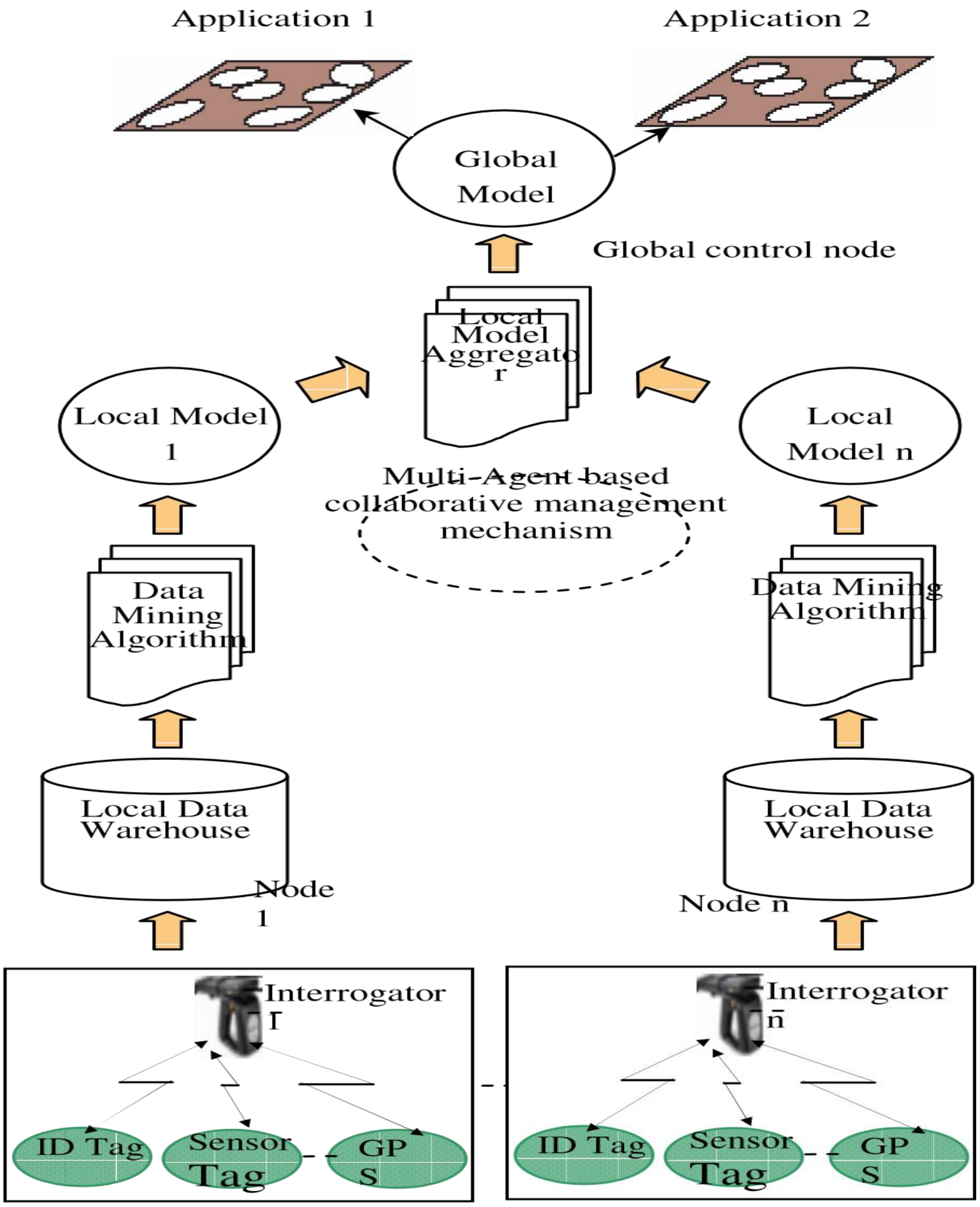

Fig 2 Distributed data model

Distributed data mining model for IoT is not only able to solve the problems brought by distributed storage of nodes, but also decompose the complex problems into simple ones. It chooses the data mining algorithm and the data sets for mining, and then navigates to the sub-nodes containing these data sets. The sub-nodes receive the raw data from various smart objects. These raw data is pre-processed by data filter, data abstraction and data compression, and then is saved in the local data warehouse. Local models are obtained by event filtering, complex event detection and data mining in local nodes. According to the demand of the global control node, these local models are submitted to the global control node and aggregated together to form the global model. Sub-nodes exchange object data, process data and knowledge with each other. The whole process is controlled by the multiagent based collaborative management mechanism as shown below in Fig 2[1].

TABLE 3
DISTRIBUTED DATA MODEL FEASIBILITY SUPPORT

TOOL	Pros by Model	Cons by Model
Weka	Weka4WS extends Weka to support remote execution of Weka data mining algorithm. To enable remote invocation, all data mining algorithms provided by Weka library are exposed as grid services.Weka4WS has been designed and developed by using the emerging Web Services Resource Framework (WSRF).This can fit in distributed data mining model of IoT at both Global node and local nodes.	Weka4WS limits to the algorithms provided by Weka.Weka4WS does not address service composition, which allows system to use services together in order to increase its usability performance.
KEEL	Primarily supporting evolutionary algorithms it could fit in local nodes alone.	More an evolutionary algorithms, fuzzy systems centric tool.
R	Purely statistical, less on data mining algorithms. It can fit in local node alone.	Purely statistical, less on data mining.
KNIME	Scalability, Intuitive user interface, High extensibility well-defined API for plugin extensions.A workflow in KNIME consists of several nodes belonging to various categories (readers, manipulators, learners, predictors, writers), which are connected via ports. A connection can either transfer data or generated models, which describe extracted information from the input data such as learned predictors or models. KNIME offers a wide array of prebuilt nodes for the execution of a multitude of different tasks. This makes it conveniently suitable for distributed data model for IoT.	The creation of workflows requires more user input and therefore is not as straightforward as local systems such as KNIME.
RAPIDMINER	Features like stream mining, Model evaluation, scripting, logging, Data partitioning, it supports multi-layer model for IoT very well.	Though it handles varied data formats like pdfs, XMLs, images, it still requires prominent knowledge of database handling.
ORANGE	Orange4WS focuses on SOAP web services with WSDL descriptions. With web service support, it can possibly fit in both global and local nodes in the model. Programming is done by placing widgets on canvas and connecting their input and outputs.	Annotations of web services and widgets are done manually, needs certain level of KD ontology and components internals. Number of widgets seem limited when compared to other tools such as RapidMiner or KNIME.

C. Grid based data mining model for IoT

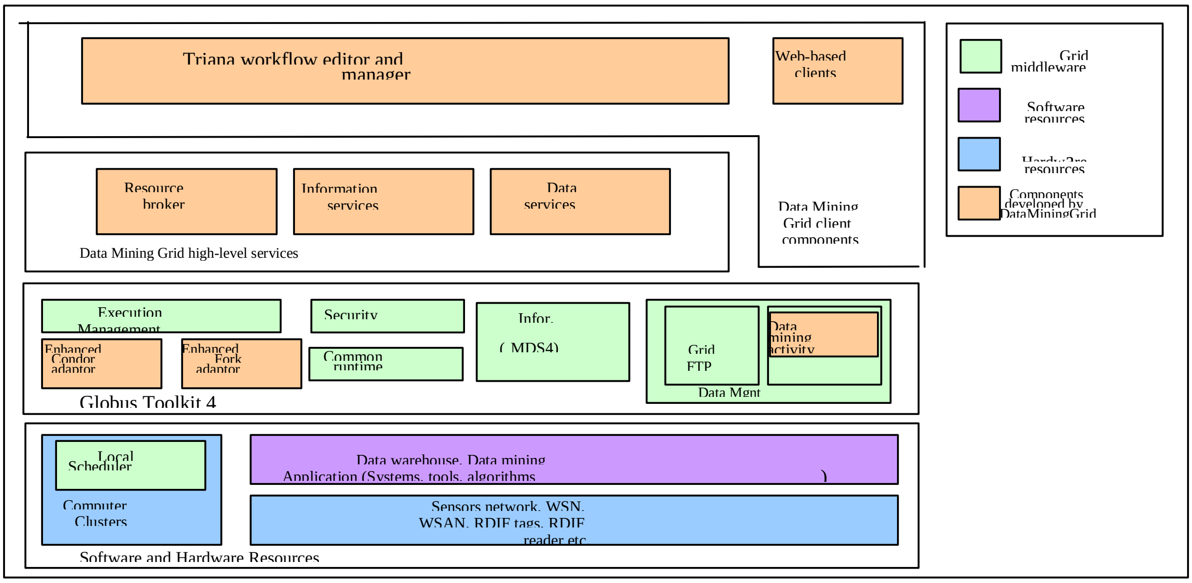

Fig 3 Grid Based Data mining model

Grid computing is a novel computing infrastructure, which is able to implement heterogeneous, large scale and high performance applications high performance applications. The basic idea of Grid is that users can make use of the computation resources of Grid as the same as power resources. It also offers various software resources, e.g., event processing algorithms, data warehouse and data mining applications etc as shown below in Fig 3[1].

TABLE 4
GRID BASED DATA MINING MODEL FEASIBILITY SUPPORT

TOOL	Pros by Model	Cons by Model
Weka	Weka4WS extends Weka to support remote execution of Weka data mining algorithm. To enable remote invocation, all data mining algorithms provided by Weka library are exposed as grid services.Weka4WS has been designed and developed by using the emerging Web Services Resource Framework (WSRF). Weka4WS, a framework that extends the Weka toolkit for supporting distributed data mining on Grid environments. This can fit in distributed data mining model of IoT.	Weka4WS limits to the algorithms provided by Weka. Weka4WS does not address service composition, which allows system to use services together in order to increase its usability performance.
KEEL	Primarily supporting evolutionary algorithms it could fit in local nodes alone.	More an evolutionary algorithms, fuzzy systems centric tool.
R	Purely statistical, less on data mining. It can complement the software resources stack to limited capability.	Purely statistical, less on data mining.
KNIME	The grid and cloud User Support Environment (gUSE) was specifically created to use distributed computing infrastructures (DCI). Scalability, Intuitive user interface, High extensibility well-defined API for plugin extension makes it convenient to be split functional and work in grid model.	The creation of workflows requires more user input and therefore is not as straightforward as local systems such as KNIME.
RAPIDMINER	Rapidminer has effective process control and has been heavily extended. It is written in Java and is platform independent and Weka framework was completely integrated into Rapid miner. It can scale well for grid based model.	Implementation of specific wrappers and checkpointing mechanisms is time consuming and huge conceptual effort for each analytical process.
ORANGE	Orange4WS focuses on SOAP web services with WSDL descriptions. Programming is done by placing widgets on canvas and connecting their input and outputs. Being component based, it can fit into Grid based model seamlessly.	Annotations of web services and widgets are done manually, needs certain level of KD ontology and components internals. Number of widgets seem limited when compared to other tools such as RapidMiner or KNIME.

IV. CONCLUSION AND FUTURE SCOPE

IoT is vast collection of inter-dependent and independent systems working in parallel. Data mining forms the crux of identifying the useful knowledge of the whole Internet eco-system and share and build it effectively. Each data mining tool can scale well in one scenario and underperform as well, it’s important to identify model and corresponding tool suite that could seamlessly be part of IoT as a whole. Research in fields of IoT and data mining can guide the tools development and performance in time. With web services, grid and cloud support it is possible to scale the existing tools, still hidden application and domain specific issues will remain yet to be explored on real grounds.

REFERENCES

[1] Rangra et al., International Journal of Advanced Research in Computer Science and Software Engineering 4(6), June - 2014, pp. 216-223 © 2014, IJARCSSE All Rights Reserved Page | 217

[2] Comparative Study of Data Mining Tools, Volume 4, Issue 6, June 2014 ISSN: 2277 128X

[3] International Journal of Advanced Research in Computer Science and Software Engineering.

[4] N.Deepika, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.1, January-2016, pg. 233-240, http://www.ijcsmc.com/

[5] http://www.webopedia.com/TERM/I/internet_of_things.html

[6] http://www.internetsociety.org/doc/iot-overview

[7] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[8] Wikipedia Encyclopedia. for IOT & Data mining.

[9] http://www.nicoschlitter.de/downloads/Distributed_Data_Analysis_using_BOINC_and_RapidMiner

[10] http://www.rapid-i.com/downloads/brochures/RapidMiner_Fact_Sheet.pdf

[11] http://ceur-ws.org/Vol-993/paper9.pdf

[12] http://www.salleurl.edu/GRSI/docs/keel_softcomputing.pdf

[13] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.454.7782&rep=rep1&type=pdf

[14] https://kops.uni-konstanz.de/bitstream/handle/123456789/5542/Parallel_and_Distributed_Data_Pipelining_with_KNIME.pdf?sequence=1

[15] http://slais.ijs.si/theses/2013-03-25-Podpecan.pdf

A.Madhuri, M.Tech student, having 5.5 years of IT Industry experience worked in Microsoft .Net framework and Microsoft Business Intelligence Studio. Currently pursuing M.Tech from New Horizon College of Engineering, VTU Belgaum.

N.Deepika, Sr.Asst.Professor having 14 years of experience in Academics has pursued her M.Tech from JNTU, Hyderabad. She is currently working In NHCE, Dept of CSE, Bangalore. She has guided many UG & PG students for their Projects. Her Research areas include Clustering techniques, Data Mining, Web Mining and Big Data Analysis.