Process of Data Visualization: Voyage from Data to Knowledge

Kirti Nilesh Mahajan[1*], Leena Ajay Gokhale[2]

[1, 2]Department of Information Technology, Institute of Management and Entrepreneurship Development (IMED), Bharati Vidyapeeth (Deemed To Be) University, Pune, India

Corresponding Author: kirtimahajan75@gmail.com,   Tel.: (91) 9823618237

Abstract—The voyage of data begins with collection of data, followed by storing this data in precise format, further traveling through the process of Data Analysis and Data Visualization, ultimately concluding the journey by reaching to valuable knowledge. Thus, this journey originates from Data, reaching to Knowledge. This original data has various dimensions, several logical formats like text, numbers and also physical structures such as structured, semi-structured and unstructured. The complexity of the data increases with the increased number of dimensions of data. During this entire journey, the data has to travel through various phases including Data Analysis and Data Visualization. However, the outcome of data analysis may not be adequate to provide the knowledge. Visualization Process involves seven steps, Acquire, Parse, Filter, Mine, Represent, Refine and Interact. Acquiring refers to obtaining the data, Parsing structures the data, Filtering allows to select the precise data, Mining supports in uncovering the patterns, Representing provides visual data, Refining allows to enhance the presentation of data, Interacting develops an interaction with the gained knowledge. Each phase makes data more meaningful as the process of data visualization contributes in enhancing the quality of the analysed data. Thus, role of visualization in this voyage is significant as it transforms data into knowledge. The purpose of this research paper is to describe the various phases in the process of data visualization along with several formats of original data and also presents comparison between data and information, before and after visualization.

Keywords— Data Visualization, Data Transformation, Visualization Process, Information.

I.   INTRODUCTION

The data is travelling through various phases of its life. The original content collected is in the raw format which means the content may be in structured, semi-structured or unstructured format. This raw data is worthless and unusable unless it is analyzed and visualized. Before analyzing the data, initially, it has to be stored in a precise structured format. Since the outcome of analyzed data may also be difficult to understand and may not convey the appropriate meaning, further performing process of data visualization is essential. Considering the gap between the original data and graphical information, converting the analyzed data into more understandable format is the key factor of gaining knowledge. Analyzing the data may not be adequate as the result of data analysis may not always generate the desired result. Due to the large size of data, the analyzed data may not be clear to understand unless it is visualized. [3] Data Visualization is essential for extracting the hidden insights from the data. The graphical presentation gives a clear picture of the facts. [2]

The modern technology give robust support in collecting the data from various sources. The outcome of this development is the large dataset with high level of variety. [6] The quantity of data captured and stored in increasing the complexity in data analysis. [7]

Information and Communication Technology (ICT) plays a vital role in the entire system of data. Internet facility and networking technology supports in data collection activity. Sharing data through networking technologies is quick and easy, Content on the websites can be easily downloaded, data is collected through online form submissions, Scanners transform the hard copy documents into digital data at a rapid pace, making multiple copies of the digital data rapidly, capturing the events or documents with digital camera or smart phone at a single click, are the key reasons of the high volume of digital data.
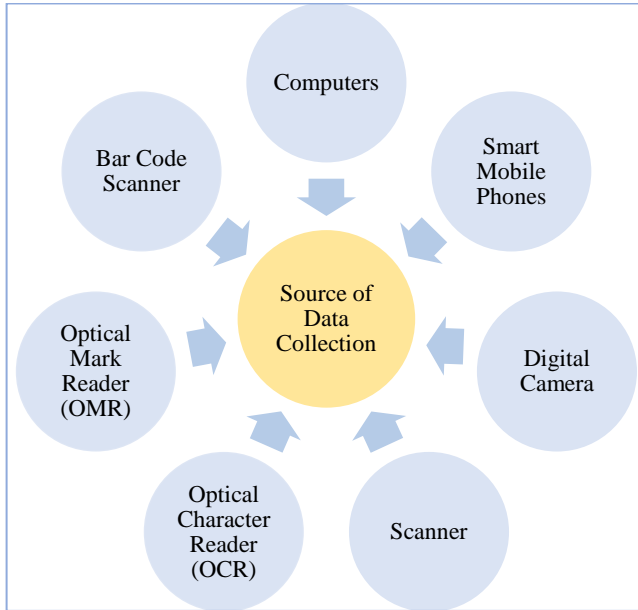
Figure 1. Several Sources of Data Collection

Hence, after the data analysis, the visualization process is carried out so as to extract and present the meaningful information from the large dataset. First phase is to 'Acquire' the data means to obtain the data. Second phase is Parse means in which the data is structured. Third phase of 'Filtering' the data is meant for identifying the requirement. Fourth phase represents Mining, which helps in uncovering the various patterns from the data. In the fifth phase, 'Represent', data is represented in the graphical format. During sixth phase of visualization, 'Refine', the representation of data is enhanced. The final seventh phase, 'Interact' allows the user to interact with the created graphic. [4]

Since the decision making entirely depends upon the information presented to the decision makers, it is essential to convert the raw and huge data into precise and compact information. This information gives knowledge which support in data driven decision making. Thus, Visualization plays a vital role in supporting the decision makers in the organizations to fulfil their information requirement.
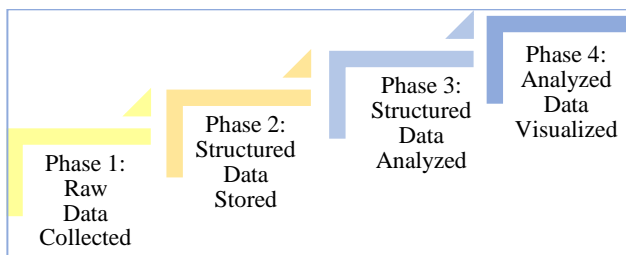

Figure 2. A Voyage from Data to Knowledge

The purpose of carrying out this study is to describe the phases of visualization process and present the comparison between data before visualization and information after visualization.

The pattern of the paper incorporates, Section I contains the introduction of voyage from original data to knowledge. Section II encompass the related work describing the significance of visualization in enhancing business decisions. Section III carries the research methodology. Section IV contains the several features of original data forms before visualization. Section V conveys various phases in visualization process, Section VI presents the comparison of Data before visualization and Information after visualization, Section VII comprises the results and discussion, Section VIII covers conclusion and future scope.

II. RELATED WORK

Various formats of data such as Structured, Semi-structured and Unstructured data are described along with their respective illustrations in the research paper titled, "Advanced Charting Techniques of Microsoft Excel 2016 Aiming Visualization". [3] The research paper, "Significance of Digital Data Visualization Tools in Big Data Analysis for Business Decisions" states the significance of visualization in enhancing business decisions. [2] In the research paper, "Comparative Study of Static and Interactive Visualization", the authors present the Key Phases in Static and Interactive Visualization Process. [1] The seven phases in the process of visualization are described in the research paper, "vizLib: Using The Seven Stages of Visualization to Explore Population Trends and Processes in Local Authority Research". [4]

III. RESEARCH METHODOLOGY

Literature review helped in understanding the concept of raw data, information and various steps involved in the process of data visualization. The purpose of this study was to convey the various phases in the visualization process.

IV. FEATURES OF ORIGINAL DATA

The original data collected has to be converted in the precise form. When the concern is about manipulating data, versatile requirements may arise. Several methods are available for manipulating the data, such as applying different functions. Sometimes, the requirement may arise when the number is to be treated as text and the text is to be treated as number. [5]
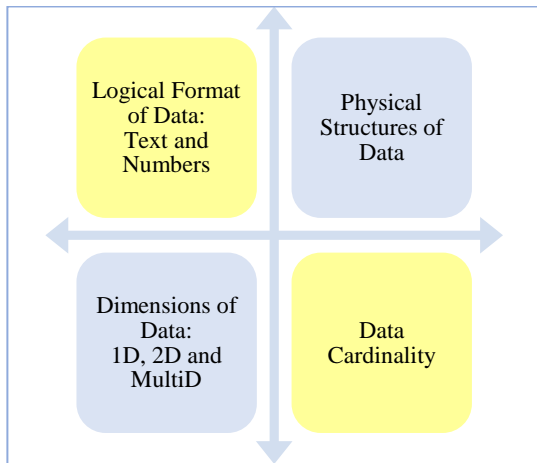
Figure 3. Features of Raw Data

A. Logical formats of Data: Text and Numbers

The data stored in a structured format may be Number Data or Text Data. Determining the type of data stored in the table or cells in a worksheet is essential. The data may be character, special character or a number. [5]

• Text Data

Different database systems may support different character set. For example, Microsoft Excel worksheet supports the American National Standards Institute (ANSI) character set and an extended character set is known as Unicode. ANSI integrates 255 characters which are numbered from 1 to 255. While ANSI character takes one byte for storage, in Unicode, each character need two bytes of storage. [5] The text data can be sorted in an alphabetical order A to Z or Z to A for better understanding and clarity of data.
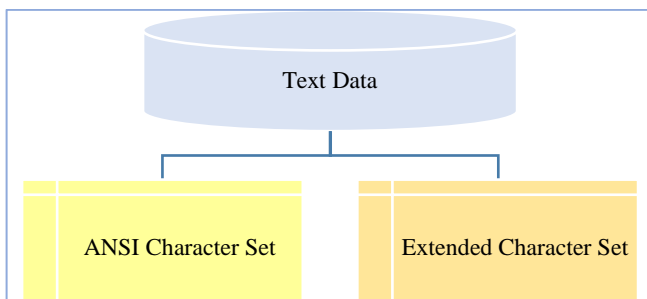

Figure 4. Classification of Text Data

• Number Data

The numbers that need some arithmetic operations can be distinguished with the numbers that do not require the calculations. Credit card number, mobile number, employee number are the examples of the number data which do not require arithmetic calculations to be performed. Hence, such number data can be forced to be treated as text by storing

number as text. Product price, profit gained, product quantity are the examples of the number data that need precise arithmetic calculations for data analysis. This calculations can be performed with the help of precise functions. Data analysis of numbers may be performed differently than the character data. [5] Both types of number data, irrespective of whether the mathematical calculations are required or not, can be sorted in ascending or descending order for enhanced understanding.
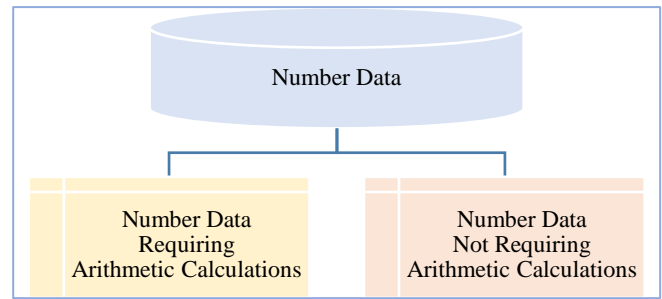

Figure 5. Classification of Number Data

B. Physical Structures of Data

The data may be classified as Structured, Semi-Structured and Unstructured. The varied sources of data collection is the key reason of this varied formats of data. The structured data refers to the data stored in precise formats in table having rows and columns, in Microsoft Excel worksheets. Data captured through online forms is also an example of structured type of data. The semi-structured data deals with data captured through web applications. Data captured through emails, social media posts can be considered as unstructured data. [3]
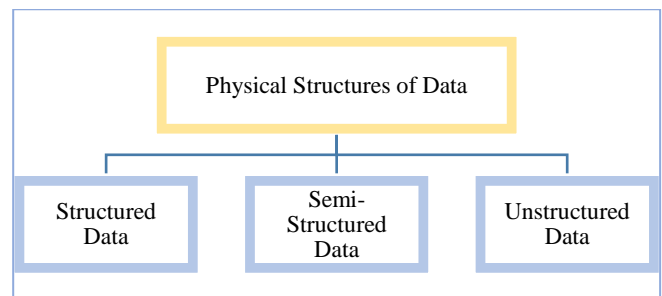

Figure 6. Physical Structures of Data

C. Dimensions of Data

As the number of dimensions in a dataset increase, the complexity of data also increases. The data having various dimensions may be visualized differently. Different chart type is suitable for different dimensional data.
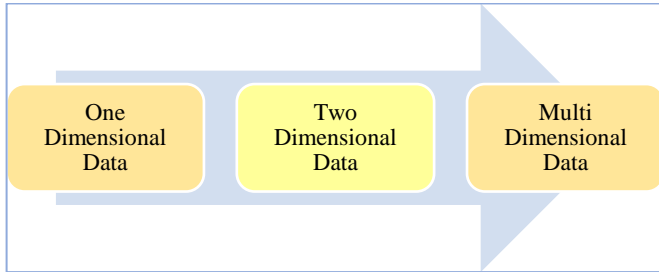
Figure 7. Dimensions of Data

• One Dimensional Data

One dimensional data has only one variable. It does not have any associated variable. This data may be Ordinal or Nominal type of Data. One-dimensional data can be represented as a worksheet having data stored in one column (one variable) and multiple rows.
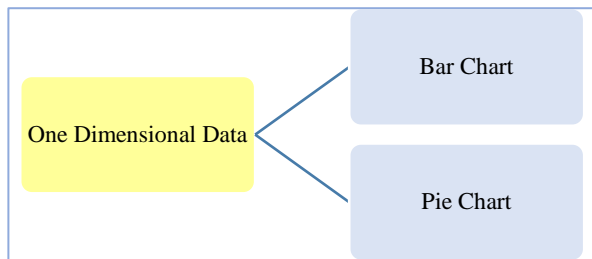

Figure 8. Suitable Chart for One Dimensional Data

• Two Dimensional Data

Two dimensional data have two variables which are associated with each other. Two-dimensional data can be represented by X-Y plots.
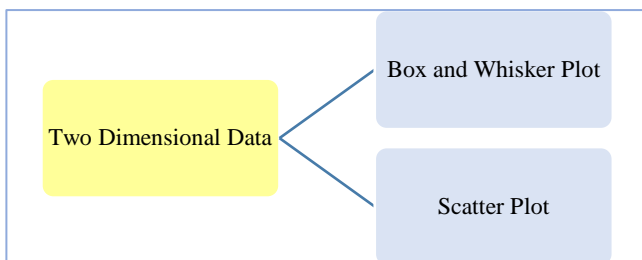

Figure 9. Suitable Chart for Two Dimensional Data

• Multi Dimensional Data

Multi-dimensional data has three or more than three variables. The association among the variables can be presented diagrammatically which reduces the complexity raised due to increased number of dimensions.
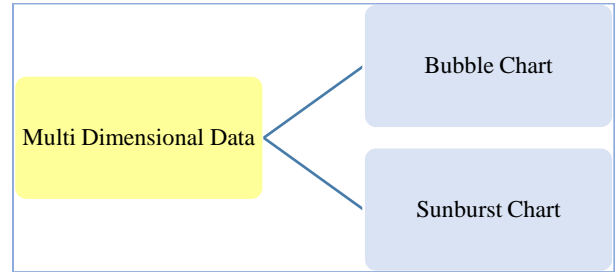

Figure 10.   Suitable Chart for Multi Dimensional Data

D. Data Cardinality

Data Cardinality is the uniqueness of data values contained in a column. High cardinality means there is a large percentage of unique values. Example product code, because each data item should be unique. Low cardinality means a column of data contains a large percentage of repeat values, such as product colour, product weight. In his case, there can be multiple records having identical values.
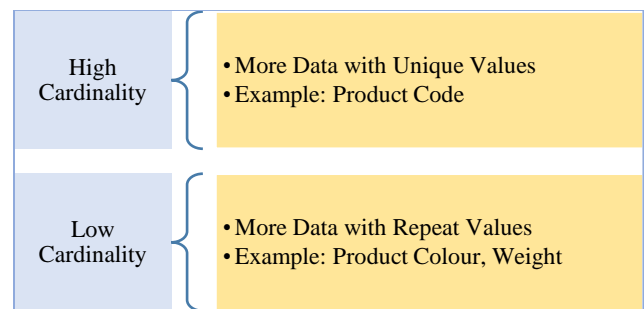
Figure 11. Cardinality of Data

V. VISUALIZATION PROCESS

The data after analysis may not always be easy to understand and may not always deliver precise meaning. For the purpose of transforming analysed data into meaningful information in the graphical format, the data has to undergo the process of visualization.

The process of visualization involves seven steps, Acquire, Parse, Filter, Mine, Represent, Refine and Interact. [4] The sequence of these phases is vital and each phase is equally significant.

Acquiring the Data

Parsing the Data

Filtering the Data

Mining the Data

Representing the Data

Refining the Data

Interacting with the Data

Figure 12.　　Seven Phases in the Process of Data Visualization

A. Acquiring the Data

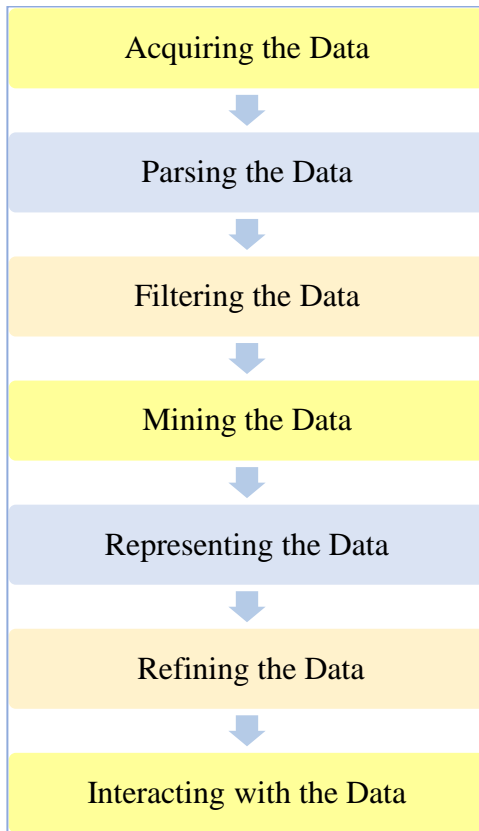This phase refers to obtaining the data from either a file on a disk or a source over a network. Attaining the accurate data is a big challenge since the data is collected from various sources. This data may in varied formats like structured, semi-structured and unstructured.

B. Parsing the Data:

Parsing the data deals with providing a precise structure that will convey the meaning of the data and ordering it into categories. Analyzing the data is the next phase which is is essential for the purpose of understanding the data. The data may be turned into a structured format by applying categorization. The data in a structured format becomes expedient to visualize.

C. Filtering the Data

Filtering refers to identifying the data of interest. [5] The data supports in enhancing the visualization by removing the unwanted data. Thus, this phase refers to extracting the relevant data for further processing by removing the surplus data.

D. Mining the Data

Mining refers to building the charts based on the data stored in a structured format. However, this step supports in getting the fundamental understanding of the data before moving ahead to the representation phase.

E. Representing the Data

The data has to be represented in a graphical format. Hence, selecting an appropriate visual type, such as a bar graph, list, or tree is essential.

F. Refining the Data

Refining the visuals is referred as enhancing the original representation for making it richer and more visually appealing.

G. Interacting with the Data

Further, interaction between the user and the charts is vital. This interaction can be made by adding several approaches for manipulating the data or governing the features to be edited.

VI. DATA: BEFORE AND AFTER VISUALIZATION

Visualization is a technique of creating graphs, charts or diagrams for the purpose of conveying a message. The transformation of data into information is the key outcome of data visualization. The comparison of data and information state clear differences between them in terms of their purpose, role, usefulness, quality and formats.

Table 1. Comparison of Data and Information

	Data	**Information**
Factor for Comparison:	*Before Visualization*	*After Visualization*
Concept:	Data is simple text and numbers.	Information is processed and interpreted data.
Purpose:	The purpose of data collection is to create the records.	The purpose of information is gaining Knowledge.
Role:	Data is not useful for decision making.	Information is useful for decision making.
Use:	Data works as a raw material for generating the information.	If the information is presented completely, accurately and on time, it

	Conclusions cannot be made without the large dataset.	helps in gaining new insights and supports business decisions.
Size:	Data is large in size.	Information can be presented in a very compact way.
Format:	Data may be in structured or unstructured format.	Precise Graphical format such as Charts.
Meaningful:	Data may not be always clearly understood and meaningful.	Visualized information is more meaningful, precise and it gives clear understanding.
Quality:	Quality of the data may be poor in order to extract the meaning out of it.	Information has quality, thus, hidden insights are presented.
Presentation:	Data may be presented in paragraph or Table Format.	Information is presented in the Graphical Format.
Origin:	Data is collected from various sources such as websites.	Graphical information is created on the basis of Data stored in a structured format.
Volume:	Data may be in large volume.	Information may be presented in a compact manner. However, a restriction should be followed on how much quantity of data to be visualized in one chart.
Time	Understanding data may require long time duration.	Information can be understood in short time duration.
Example:	Text data, Numeric data, Alphanumeric Data	Graphical presentation like Charts

VII. RESULTS AND DISCUSSION

The journey of data has various stages such as Collection of data, Storage of data, Analysis of data and Visualization of data. At the start point of collection, the data is in the assorted format. Data becomes ready for analysis after completing the storage step in which it is stored in the uniform layout. The outcome of Data analysis helps to understand the complex data. However, this data may not be useful until it delivers the precise meaning out of it as data analysis may not be sufficient and suitable to extract meaningful information. Visualization process transforms analysed data into information displayed in graphical pattern like charts. Graphical presentation of data helps in accurate decision making by delivering hidden insights. During visualization process, the data travels through seven key steps, Acquire, Parse, Filter, Mine, Represent, Refine and Interact. As the data passes through each of these phase, it gets enhanced. This journey of data reaches to the destination of important knowledge gaining.

VIII. CONCLUSION AND FUTURE SCOPE

This paper presents the process of data visualization which transforms the raw data into the meaningful information. Considering the gap between the raw data and graphical information, visualization plays a vital role in supporting the decision makers in the organizations to fulfil their information requirement. Since the decision making entirely depends upon the information presented to the decision makers, it is essential to convert the raw and huge data into compact and precise information.

ACKNOWLEDGMENT

REFERENCES

[1] Kirti Nilesh Mahajan and Leena Ajay Gokhale, "Comparative Study of Static and Interactive Visualization Approaches", International Journal on Computer Science and Engineering (IJCSE), e-ISSN: 0975-3397 p-ISSN: 2229-5631, Vol. 10 No.03, DOI: 10.21817/ijcse/2018/v10i3/181003016 Vol. 10 No.03 Mar 2018 85, pp.85-91, March 2018.
[2] Kirti Mahajan and Leena Ajay Gokhale, "Significance of Digital Data Visualization Tools in Big Data Analysis for Business Decisions", International Journal of Computer Applications (IJCA), Volume 165 – No.5, pp.15-18, May 2017.
[3] Kirti Mahajan and Leena Ajay Gokhale, "Advanced Charting Techniques of Microsoft Excel 2016 Aiming Visualization", International Journal of Computer Sciences and Engineering (IJCSE), Vol.-7, Issue-1, E-ISSN: 2347-2693, Jan 2019.
[4] Robert Radburn, Jason Dykes, Jo Wood, "vizLib: Using The Seven Stages of Visualization to Explore Population Trends and Processes in Local Authority Research".
[5] Michael Alexander and Dick Kusleika, "Excel 2016 Formulas", Wiley India Pvt. Ltd., India, pp. 99-123, 2016.
[6] Mantripatjit Kaur, Anjum Mohd Aslam, "Big Data Analytics on IOT: Challenges, Open Research Issues and Tools", International

Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.81-85, 2018.

[7] Rakesh. S.Shirsath, Vaibhav A.Desale, Amol. D.Potgantwar, "Big Data Analytical Architecture for Real-Time Applications", International Journal of Scientific Research in Network Security and Communication, Volume-5, Issue-4, ISSN: 2321-3256, 2017.

Authors Profile

Kirti Nilesh Mahajan pursed PhD from Bharati Vidyapeeth University, Pune in year 2012 and Master of Computer Applications from Gandhi University, Rajasthan in year 2004. She has pursued Advanced Program In Human Resource Management from IIM Lucknow in year 2014, Java Programming & Java Scripting Certification from APTECH, Mumbai in year 2001. She is currently working as Head of Department, Information Technology. She is working as Professor in Institute of Management & Entrepreneurship Development, Bharati Vidyapeeth Deemed to be University, Pune, INDIA since 2012 and also as Director of International Students' Center since 2016, Director of Department of Students' Welfare since 2014, Joint Director of Academic Staff College, Bharati Bhawan, Pune, since 2013, Coordinator of European Union Research Collaboration & Funding Programs with Bharati Vidyapeeth (Deemed to be) University, Pune, India since 2013. She had been working as Associate Professor from 2005 to 2012. She has won many awards including "International Achievers Award for Education Excellence" by International Achievers Foundation, New Delhi, 25th January 2014, Bangkok, Thailand. She has total 20 years' of experience as an Analyst and Researcher, Professor. She is a member of Editor Advisory Board, International Journal of Research & Development (ISSN Online 2279 073X & ISSN Print 2279 0438) and various othe committees including Research Development Committee, IMED, BVDU, Pune. She has published more than 25 research papers in reputed international journals including International Journal of Computer Applications (IJCA) and International Journal on Computer Science and Engineering (IJCSE).

Leena Ajay Gokhale pursed Master of Philosophy (IT) from Bharati Vidyapeeth Deemed University, Pune, India in 2010, Master in Computer Management from Pune University, Pune, India in the year 1997, Diploma in Computer Applications from Symbiosis Institute of Computer Science and Research, Pune, India in 1995. She is currently pursuing Ph.D. and working as a Visiting Faculty in Institute of Management & Entrepreneurship Development, Bharati Vidyapeeth (Deemed to be University), Pune, India since 2007. She has published more than 10 research papers in the international and national conferences and reputed international journals including International Journal on Computer Science and Engineering (IJCSE) and International Journal of Computer Applications (IJCA). Her main research work focuses on Visualization Technology. She has worked as a Visiting Faculty with renowned institutes in Pune. She has a teaching experience of 12 years.