An India Towards Global Ranking Standards: A Move Towards Data Curation

Kashinath Chandelkar, Piyush Gupta

Abstract*: Ministry of Human Resource Development (MHRD) is not only a source of inspiration for sectors such as Research and Development, Services, Agriculture and Entrepreneurship. They are also enhancing our competitiveness and relevance by proposing and executing different projects across the nation. The approach is in line with the vision 2030. As per 2018 Shanghai report of Times Higher Education (THE), only 40% academic and 10% of employer reputation is built as per the Global Ranking Standards. Indian Higher Educational warehouses with the help of National Institutional Ranking Framework (NIRF), Rashtriya Uchchatar Shiksha Abhiyan (RUSA) and University Grant Commission (UGC) are generating data which is of least relevance to end users. We propose peer to peer data curation model which shall be extended as per business need using a cloud technology. We have harnessed data related to the Higher Education Sector (HES) from Indian Institute of Sciences (IISc), RUSA, and MHRD, to understand Indian contribution towards global ranking standards in Higher Education Sector. The approach is in line with the Escuela Nueva Model and Open Architecture Information System (OAIS) as existing standards in traditional data curation process. It is supported by the patent in large-scale data curation.

***Keywords* : THE, UGC, RUSA, OAIS, MHRD, NIRF, HES,IISc,Data Curation.**

I. INTRODUCTION

Content created by the Higher Education Sector in India is backed by a quality learning environment [1] delivering quality content to students and other authorized end users. Our institutions have a best possible technological support in terms of content creation and management for data curation practices. The question still arises why Indian institutions [21] are not among the top 100 universities? The probable reason maybe we are not in line with global standards [3] [16] or we have diverted from the mission 2030 [8]. Even after having a potential to contribute globally in manufacturing, services, human resource and agriculture, we are ranked 179th in ease of doing business. We may also have areas for improvement across the educational system. Each institution strives hard towards growth, by prioritizing the daily needs of an individual and organization as a whole. The growth is affected by the outcome generated from within the institution and governing bodies. Data generated over the period of time in its different formats is an asset to the institution. Making this asset available to end user is a challenge. Data loss during its processing cycle or digitization is another dimension to explore.

Recently, an intruder had an access to 87 million Facebook users [4] name, address, email id, and recent history. This incidence indicates data in public and private domain is not preserved properly. This motivates to identify a solution that preserves trusted data in the Indian higher education sector for efficient decision making. The probable solution lies in data curation which is keeping trusted data secure over its processing cycle and making it available to end user in a suitable file format. How to identify trusted data from huge data corpus irrespective of language and file formats is an area of concern.

For a given data corpus, extracting information worth decision making in a given frame of time is a challenge. Traditional methodologies have also made an attempt and have succeeded to show its presence in data storage and retrieval practices. Data loss and exposure is seen while selecting corpus, processing it using contemporary practices and displaying relevant information. Our peer to peer data curation design helps to identify trusted data through the curation process. It is tested for three different data corpus to identify how India is contributing in line with the global ranking standards. Identification of trusted data shall minimize the cost, space, and efforts required to store, process and retrieve data over the network. This content is shared and preserved as per Open Architecture Information System.

II. RELATED WORK

The reviewed literature is broadly categorized into Indian standards, Global ranking standards and Data curation Practices.

2.1 Indian Ranking Standards

New trends in education system [14] are analyzed with the help of google. We can interact with the system even after having its limitations with the help of query. Keeping in mind vision 2030 [8] nation require the skilled labor of 75 million peoples in the field of Agriculture, Manufacturing, and services. With existing 40% contribution in academics and 10% in employer, to fulfill global demand. National Institutional Ranking Framework(NIRF) [12] have taken multiple initiatives to work towards global ranking standards. The details are summarized in Table-1. The Indian government believe that all can

Revised Manuscript Received on January 05, 2020
***** Correspondence Author
Kashinath Chandelkar, Birla Institute of Technology Mesra, Department of Computer Science & Engineering, Jaipur Campus,
Dr. Piyush Gupta, Birla Institute of Technology Mesra, Department of Computer Science & Engineering, Jaipur Campus,

learn if learning environment and quality content are delivered to the right user. The Escuela Nueva model [1] is in favor thereby creating a student-centered process for service delivery.

Table-1: Indian Ranking Parameters summary and Weightage

Parameters	Marks	Weightage
Teaching, Learning & Resources	100	0.30
Research and Professional Practices	100	0.30
Graduation Outcomes	100	0.20
Outreach and Inclusivity	100	0.10
Perception	100	0.10

2.2 Global Ranking Standards

The Global Ranking Standards [15] are into existence from 2003. The Academic Ranking of World Universities (ARWU), best Known as Shanghai Ranking has six indicators, broadly categorized into a quality of Education, Faculty, Research output, and Per Capita performance. The Quacquarelli Symonds (QS) world university ranking is evaluated based on six matrices which include academic and employer reputation, faculty-student ratio, citations per faculty, followed by international faculty and student relation. The Times Higher Education [3] contributes in global ranking by using different performance indicators as shown in Table-2.

2.3 Data Curation

Maintaining the value of digital data over its processing cycle is termed as data curation [5]. OAIS Model [9] helps in making data available in all supported file formats. Digital Information Migration [10] is needed in the process of maintaining information integrity and security caused due to legal and institutional issues. Traditional data in the existing data corpus is designed for programmers and is of least importance to an end user. Extracting data from the data source and ingesting it into the curation process is time-consuming. Complexity increases if the data is noisy [2] or in different file formats. Traditional algorithms [23] are widely used in the process. Algorithms cover data extraction using indexing, querying and ranked retrieval [20]. [7] Clustering helps in summarizing data from different file formats. Digital Curation Center (DCC) suggests the latest trends.

III. PROPOSED METHODOLOGY

In the case of data curation, following patents [6] have an important contribution to this design. US Patent No: 6567814 METHODS AND APPARATUS FOR KNOWLEDGE DISCOVERY IN DATABASE (Bankier, May 20, 2003). discloses in the abstract " project plan is created using Graphical User Interface. An object in the project represents a functional component. Links within the objects represent a flow of data from source to sink. Data visualization component can be inserted anywhere in the project plan. Compression technology is applied to reduce the overall size of a database." US Patent No: 20100179930 Ai Methods And Systems For Developing Prediction From Desperate Data Sources Using Intelligent Processing (Teller At Al July 15, 2010) Discloses In The Abstract " Prediction Is Done Based On The Extracted Features And Observations Collected From Disparate Data Sources. They Have Focused On Feature Selection For Specific Predictions." Pct Patent No: W0201401257 A1 A Method And System For Integrating Data Into Database (Beskales At Al, Jan 23, 2014) Disclose In The Abstract " About Integrating Data In Databases. The System Has A Learning Module And Duplicate Elimination Module That Works Respectively."

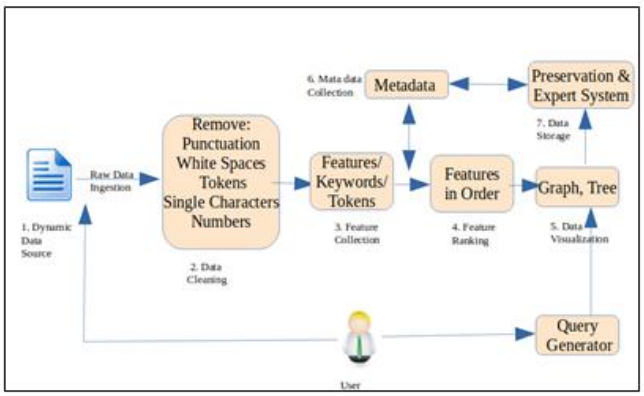

Fig-1: Data Curation process

Table-2: Global Ranking Parameters and weightage

Ranking	Indicator	Weight (%)	Weaknesses	Strengths
ARWU	Alumni	10	Old universities have more alumni strength,	Rewarded for quality work
	Award	20	Does not represent all areas, limited achievers	Shall attract international contributors for the cause
	HiCi	20	University may hire candidate from this category	High quality pool of researcher contribute in the process
	NS	20	Covers only research areas from natural science , a wide spectrum still available to explore	Identified journals represents quality output
	PUB	20	Number of publications are considered , quality of work is not on priority	Represents influence and university impact
	PCP	10	May encourage universities to limit their staff	Representative of the quality
QS	Academic Reputation	40	Already established universities may be in win win condition, lack of transparency	Good universities are given more opportunities to sustain
	Employer Reputation	10	Already established universities gains profit	Importance is given for teaching

	Faculty /Student Ratio	20	Lack quality	simple
	Citations per faculty	20	Lack transparency, citing research work within university is possible	Self citations not counted, may impact research
	International Faculty + student Ratio	5+5	Quality not on priority, good for small counties	Provides information on international attractiveness
THE	Teaching	30	Focuses on Faculty/Student ratio, institutional income, can be altered	Minimizes influence on each internal parameter
	Research	30	Focus on reputation, income and research productivity	Minimizes influence on each internal parameter
	Citations	30	Biasing may happen due to self citations and article with different author	Field weight , represents impact on research
	International Outlook	7.50	Cf(1), international collaborations can be bought, may be biased	cf(1)
	Industry Income	2.50	Depends on national policies	Low wight

The data curation architecture shown in Fig-1 is inspired by the traditional data curation process [5]. It works based on common sense to identify the root of the problem. The existing architecture speaks about collecting data from the source and ingesting into the data corpus. On processing, the data is visualized in the form of a graph with a time stamp. The available information is preserved for decision making. Current research neither speak about the content available in the data corpus, nor discloses process used for data cleaning. Since obtained keywords were not ranked it, hampers the quality of content shown to end user.

Proposed peer to peer data curation is object-oriented in nature. For each data corpus, the content is manually searched and extracted. It is categorized and clustered using traditional algorithms. The categorization continues at character level during data cleaning. The dynamic data source used for ingestion consist of varies file formats. Files having extensions .txt, .pdf, .xls, .csv and .docx are grouped under category α. Files with extensions HTML, PHP, XML, ASP and JSON are marked under group β. Audio and video related formats are categorized into δ. The remote secure connection that's work on real-time is placed in ε category.

Any system files if supported as a data source binds into ζ group.

Free and open source tools are used in a complete data curation process. R studio having Ubuntu 18.10 is used with the latest upgrades and updates. Rapid Miner has helped in data cleaning, clustering, and even visualization. Rattle with 5.4.2 accepts data corpus for analysis. Voyant tool from GitHub summarizes token for relational modeling. Metadata about keywords and extraction process is preserved to improve performance over the period of time. The data is plotted to identify words per sentence and vocabulary density followed by relative frequency for most highly ranked keywords. Visualized data is made available in widely accepted formats. The idea is tested with an applied algorithm having the pseudo code as shown in Figure-3.

IV. RESULTS & DISCUSSIONS

4.1 Data sets

As per the global university ranking standards of 2018, IISc Bangalore [11] is ranked first among the Indian institutions. A data corpus of 6.5 GB is created having content with different file formats. Basically, video, articles, and documents both online and offline from the IISc Bangalore are collected. Documents are classified and clustered for easy retrieval. The analysis is still a challenge due to the difference in file formats. Content is summarized in 7 different system readable forms after the initial cleanup, making a size of 2.3 MB.

```
READ Dynamic Data Corpus
    DETERMINE :
    IF documents are classified with similar file format
        REPEAT Document classification
        CALL Data Cleaning
        UNTIL Document<= Data Corpus
    Data Cleaning
        READ Document
        IF Document == Character String
        FOR (Document=0; Document<= Data Corpus; Document ++)
            GET Punctuation
            GET Special Characters
            GET Numerical Values
            GET Single Characters
            CALCULATE TF/IDF
            PRINT Term
            CALL  METADATA
        END FOR
        END IF
        ELSE OBTAIN Character String
    READ METADATA
        FOR (Term=0; Term<= Character String; Term ++)
            Rank Term
            CALCULATE Relative Frequency
            DISPLAY Relative Frequency
        END FOR
    END IF
```

Figure- 3: Pseudo code for Proposed Data Curation algorithm

The reliability of content is low as it has the least impact on end users. A data set from Rashtriya Uchchatar Shiksha Abhiyan (RUSA) [12] was identified making a size of 3.5 GB. A similar corpus creation process was carried out for RUSA and MHRD. The reliability of content is higher as it affects most of the Indian institutions in terms of funds for technical projects and job creations if any. After initial processing, the size was reduced to 904 KB. Similarly, corpus from MHRD is identified having a size of 7.5 GB. It consists of reports and other content for the year 2017-18. The compressed size is 44.3 KB. This corpus was highly reliable as decisions made based on these facts affects institutions across the nation in India.

4.2 Experiment & Discussion

4.2.1 IISc Results and Discussion

IISc Bangalore has widely accepted institution for its contributions in the field of research and development, its recent contributions are published in reputed journals. Since each publication has its reference to the page, PP token was ranked higher than a journal. Since it

does not make much relevance to results and end user it is neglected and the next token is considered. Fig 4.2.1 shows the contribution made by different scientists viz Kumar, wang, and Kim research work shows acceptance in IEEE journals. Being journal is the only medium for publications and knowledge transfer,is ranked highest. As the documents in corpus increases, relative frequency among the documents decreases exponentially. A material token discloses the resources provided to scientists at the institution for the said purpose and beyond.

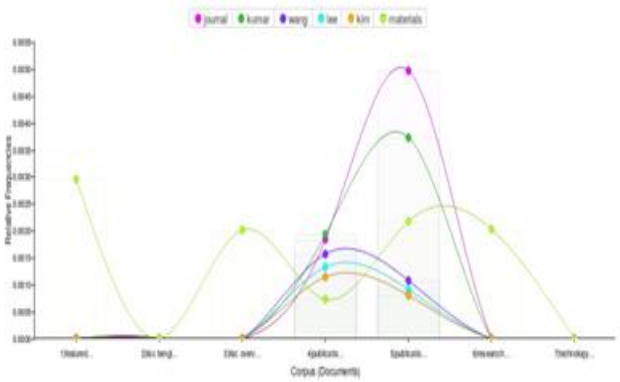

Fig-4.2.1: Relative frequency of tokens in IISc Bangalore corpus

The detailed view in corpus shows document name featured research having 169.5 words per sentence. It is calculated using sentence2vector and paragraph2vector technique. It is followed by the IISc overview document that gives an idea about different amenities given to thestaff. Technology transfer document has 21.4 words per sentence. This shows that technology transfer is still in its puberty. The schemes given to staff have 41.5 words per sentence. The data stream when evaluated at paragraph level gives vocabulary density for a given document. Density in chronological data curation for five documents is 0.542 %. In the case of peer to peer curation, 0.52 % was observed during comparison with chronological data curation.

4.2.2 RUSA Results and Discussion

Rashtriya Uchchatar Shiksha Abhiyan (RUSA) being a government of India Initiative is assumed to be in line with vision 2030 and global ranking system. The documents are government orders to the concerned authorities hence bears the higher priority.

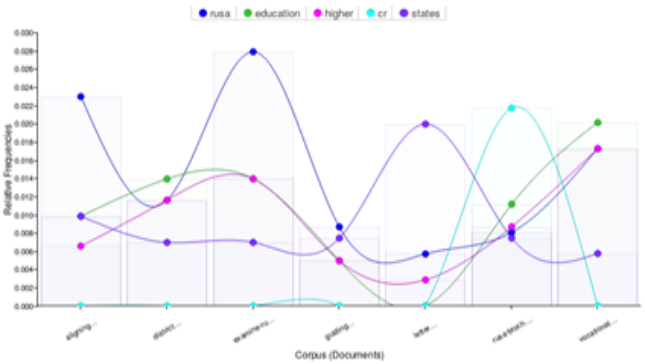

Fig-4.2.2: Relative frequency of tokens in RUSA corpus

The graph in Fig-4.2.2 speaks about higher education across the nation, the different projects initiated, completed and managed by RUSA by providing funds to the institution.

The relative frequency plotted as a part of relational modeling is based on average words per sentence streamlined from a stream of tokens after initial data cleaning. A document named aligning with RUSA shows 30.5 words per sentence, it indicates that institutions across India are working on mutual understanding to share resources among other institutions. The document named distance examination was found with 16.6 average words per sentence, this demands attention while maintaining the quality of a project. Document having principles, vocationalisation of education and alignments with RUSA have vocabulary

density of 0.74, 0.66 and 0.64 respectively, it indicates that the government is investing in skill development while focusing on principles to achieve international demand towards vision 2030. since most of the documents are extracted from the similar data source, a keyword is dominating in the relative frequency.

4.2.3 MHRD Results and Discussion

The data corpus is broadly categorized under HES Infrastructure which includes colleges, universities and available hostels premises. Teaching faculty ranked as the Professor, Associate Professor and Assistant Professors are included under the group teaching and research. Indian candidate who has enrolled for Masters, M. Phil, and Ph. D

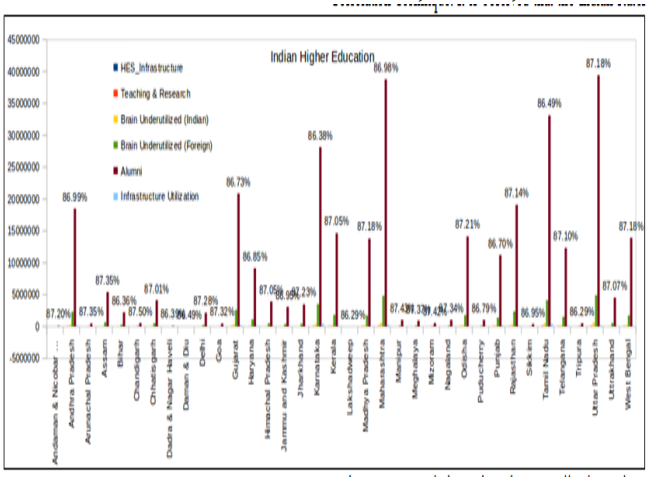

Fig-4.2.3 MHRD Data Indicating Alumni (2017-18)

and is in the process of exploring education to apply in real life categorized as Brain Underutilized Indian Group.

Similarly, foreign students enrolled in these categories are counted under Brain Underutilized Foreign Group. All the students from higher education sector who have qualified examination and are in search of a job are counted under Alumnus. These skilled candidates are the real focused area for MHRD dataset as shown in fig 4.2.3.

PSPP, the open source software, is used to analyze the numerical data set. As per the patent [6] state wise alumni are focused to understand government initiative and student contribution in a given environment. All the factors in the above-mentioned dataset are correlated using Karl Pearson Correlation Technique. It is observed that the Indian states such as Uttar Pradesh, Maharashtra, Tamilnadu, Gujarat have more Alumnus In search of a job followed by West Bengal, Rajasthan, and Andhra Pradesh.

While correlating these variables with alumnus from MHRD dataset, it is observed that faculties at different levels from the respective organization grouped under “Teaching & Research” have a major contribution in knowledge creation and transfer. Registered students for Masters, M. Phil. And Ph. D from India and Abroad termed as “Brain underutilized Indian and Brain underutilized Foreign”. These are highly correlated having value=1, on a scale of 0-1. The relation shown by correlations as “Existing infrastructure and infrastructure utilization” show the value of 0.99 and 0.86 respectively. Total grants for colleges and universities under the head planned and unplanned are correlated having a value of 0.63 and 0.19 respectively, it shows that most of the funding is made through planned procedures. Unplanned funding to the institutions may be granted for recovering in case of the institution is suffering from natural calamities or peer pressures. Considering the facts and the student's enrollment ratio in higher studies, India is found to be a nation having the potential to march towards its vision.

V. CONCLUSION

We began with the aim in mind to understand Indian contribution in global ranking standard. The Shanghai ranking system have norms in place where as India contributes with principles in place in governing bodies like UGC and RUSA. The major issue is improving Indian ranking standards as with reference to Global University ranking Standards. It was observed that in India vast amount of available data cannot be used by the end user as it is only designed for a programmer. Existing data management is done using the chronological data curation system which works on common sense having its own limitations. One of the limitations was ranking of identified tokens in Indian scenarios.Using proposed data curation model in Indian Context, tokens are identified from each data corpus. These token are ranked and visualized for decision making. It is observed from the obtained results that IISc Bangalore is contributing in terms of publications in reputed journals, these publications have helped professionals working in specialized fields to enhance research. Results available from RUSA shows that it is bridging the gap between academia and industry. Proper utilization and management of projects across the Indian states shall improve performance. MHRD have under utilized brain from India and abroad which needs proper channelization and integrating with the existing Higher education system. This shall create quality terms and intern shall help in better decision making. In India it is proposed that a hybrid utilization of available resources from higher education sector shall create better data curation process. we have only considered Korean language in the process of data curation with satisfactory results, incorporating Hindi, Sanskrit and other languages for data curation are still an open challenge as future endeavor.

REFERENCES

1. D. Adams, "Defining quality in education," Improv. Educ. Qual. Proj., vol. Biennial R, no. 1, p. 14, 2000.
2. Apostolova Emilia and A. Kreek, "Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data," Empir. Methods Nat. Lang. Process., 2018.
3. T. Barratt, "methodology for wall street journal times higher education college ranking 2019.pdf," Times Higher Education online report, 2018. [Online]. Available: www.times higher education.com/.
4. By Mike Isaac, Sheera Frenkel, "Facebook Security Breach Exposes Accounts of 50 Million Users," the new York times, 2018. [Online]. Available: https://www.nytimes. Com /2018/09/28 / technology /facebook-hack- data- breach.html. [Accessed: 28-Sep-2018].
5. S. Choi, J. Seo, M. Kim, S. Kang, and S. Han, "Chrological Big data Curation : A Study on the Enhanced Information Retrieval," IEEE Access, vol. 14, no. 8, pp. 1-9, 2016.
6. S. Clara et al., "methods and systems to train models to extract and integrate information from datasources," us 9542412 b2, 2017.
7. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," IEEE Access, vol. 5, pp. 15991-16005, 2017.
8. Ernst & Young, "Higher Education in India: Moving Towards Global Relevance and Competitiveness," FICCI, p. 76, 2014.
9. P. Gerth, "Data Curation: How and why. A showcase with re-use scenarios.," Stud. Digit. Herit., vol. 1, no. 2, p. 182, 2017.
10. L. Group, "Preserving digital information: A review," Arch. Museum Informatics, vol. 10, pp. 148-153, 1996.
11. IISc Bangalore, "globally ranked institution," IISc Bangalore, 2018. [Online]. Available: https://www.iisc.ac.in/ .
12. N. Institutional and R. Framework, "india rankings 2018 national institutional ranking framework methodology for ranking of academic institutions in india ministry of human resource development," nirf online rep., p. 30, 2018.
13. M. Jalali, A. Bouyer, B. Arasteh, and M. Moloudi, "The Effect of Cloud Computing Technology in Personalization and Education Improvements and its Challenges," Procedia - Soc. Behav. Sci., vol. 83, pp. 655-658, 2013.
14. S. Martin, E. Lopez-Martin, A. Lopez-Rey, J. Cubillo, A. Moreno-Pulido, and M. Castro, "Analysis of new technology trends in education: 2010-2015," IEEE Access, vol. 6, pp. 2010-2015, 2018.
15. K. Miller, "Steps to Research Data Readinessâ€™. DCC Briefing Papers. Edinburgh: Digital Curation Centre," digital curation conference, 2018. [Online]. Available: http://www.dcc.ac. uk/resources/ briefing-papers. [Accessed: 28-Sep-2018].
16. M. Mussard and A. P. James, "Engineering the global university rankings: Gold standards, limitations and implications," IEEE Access, vol. 6, pp. 6765-6776, 2018.
17. R. Nagpal, C. Wadhwa, M. Gupta, S. Shaikh, S. Mehta, and V. Goyal, "Extracting Fairness Policies from Legal Documents," IBM Res., pp. 1-9, 2018.
18. S. On, "All India Survey on Higher Education 2017-18," MHRD online Rep. 2018, vol. 12, p. 269, 2013.
19. RUSA, "Rashtriya Uchchatar Shiksha Abhiyan," Government of India, 2018. [Online]. Available: http://rusa.nic.in/about-us/guiding-principles/.
20. M. Sanderson and W. B. Croft, "The history of information retrieval research," Proc. IEEE, vol. 100, no. SPL CONTENT, pp. 1444-1451, 2012.
21. Singh, "why indian universities are not in top 100," online source, 2017. [Online]. Available: https://www.ndtv.com/education/world-university -ranking-why-indian-universities-are-not-in-top-100-1746663. [Accessed: 29-Sep-2018].
22. D. Wood, "Linking enterprise data," Link. Enterp. Data, pp. 1-291, 2010.
23. X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2008.

AUTHORS PROFILE

Mr. Kashinath Chandelkar is working as Project Assistant-III at CSIR- CEERI (Central Electronics Engineering Research Institute) Pilani in Smart Sensor Department, Government of India. Before Joining CEERI Pilani he was working as Corporate Trainer & Developer for CDAC (Center for Development of Advanced Computing), An organization under Ministry of IT & Telecommunication. He has 8 years of teaching experience as Assistant professor and Lecturer for both PG and UG Classes. He has earned 6 years of research experience from CEERI, ICAR KVK, CDAC and BITS (Birla Institute of Technology Mesra), Jaipur Campus as full time PhD Scholar in the department of Computer Science & Engineering. He has published 6 papers in reputed journals and 5 in international conferences to his credit.

Dr. Piyush Gupta is working as Assistant Professor in the Department of Computer Science & Engineering at Birla Institute of Technology Mesra, Jaipur Campus. He has contributed as Head & Academic Coordinator at BITS International Center Muscat (Oman). He has also served as IT Head at Oriental Bank of Commerce at Head Office, New Delhi. He has presented papers and participated in more than 25 international conferences. Dr Gupta has attended 30 workshops, seminar & conference. He has also conducted Corporate Training and workshops in India and abroad. He is a life member of Indian Accounting Association and Computer Society of India, Hyderabad. Currently two students are pursuing PhD under his guidance