Research Article

Efficient Methodology for Deep Web Data Extraction

Shilpa Deshmukh[1]*, P.P. Karde[2], V.R.Thakare[3]

[1*.]Department of Computer Application SIES college of Management Studies Navi Mumabi, India
[2.]Department of Information Technology, Government Polytechnic, Amravati, India
3.Department of Computer Science, SGB Amravati University, Amravati, India
shilpad@sies.edu.in[1], p_karde@rediffmail.com[2], vilthakare@yahoo.co.in[3]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Deep Web substance are gotten to by inquiries submitted to Web information bases and the returned information records are enwrapped in progressively created Web pages (they will be called profound Web pages in this paper). Removing organized information from profound Web pages is a difficult issue because of the fundamental mind boggling structures of such pages. As of not long ago, an enormous number of strategies have been proposed to address this issue, however every one of them have characteristic impediments since they are Web-page-programming-language subordinate. As the mainstream two-dimensional media, the substance on Web pages are constantly shown routinely for clients to peruse. This inspires us to look for an alternate path for profound Web information extraction to beat the constraints of past works by using some fascinating normal visual highlights on the profound Web pages. In this paper, a novel vision-based methodology that is Visual Based Deep Web Data Extraction (VBDWDE) Algorithm is proposed. This methodology basically uses the visual highlights on the profound Web pages to execute profound Web information extraction, including information record extraction and information thing extraction. We additionally propose another assessment measure amendment to catch the measure of human exertion expected to create wonderful extraction. Our investigations on a huge arrangement of Web information bases show that the proposed vision-based methodology is exceptionally viable for profound Web information extraction.

Keywords: Deep Web, Web Data extraction, ViDE, VBDWDE

1. Introduction

Web content mining is the way toward separating centre substance from web records. The term content extraction was presented by Rahman [1]. As the web develops quickly, anybody can transfer or download any data whenever. This prompts the ceaseless extension of insignificant, repetitive, organized and unstructured data on the site pages. A web report may contain sound, video, text, pictures, tables and so forth Extricating valuable data from these sorts of unstructured information is an intricate undertaking. A few calculations were produced for this reason and everyone has its preferences and inconveniences.

Content extraction has numerous focal points. It is simpler for getting to helpful data in a convenient and proficient way. Unessential and excess data is taken out. Since it won't burn through their time and memory for ordering and putting away unimportant substance, the presentation of web index is expanded. So it very well may be considered as a pre-processor for internet searcher. It likewise causes clients to peruse web through little screen gadgets. It additionally helps in creating rich website rundown from web journals or articles.

Typically a web content extractor, removes all the data on the site pages including text, illustrations, sound, video, joins, commercials, substance, and so forth During the extraction cycle, uproarious information are disposed of and helpful data is safeguarded. Numerous calculations were created for wiping out these boisterous data and extricating the center substance of the pages.

2. Literature Review

Numerous creators have attempted to misuse content extraction apparatuses for web reports. A few features of the significant work are sketched out here.

Sandip et al [2] proposed the programmed ID of instructive areas of pages. Here four straightforward yet incredible calculations called Content extractor, Feature Extractor, K-Feature Extractor a L-Extractor were proposed to recognize and isolate content squares from non-content squares. Highlight Extractor depends on the portrayal and utilizations heuristics dependent on the event of specific highlights to recognize content squares. K-Feature Extractor is a unique change of Feature Extractor which perform better in a wide assortment of site pages. Content Extractor distinguishes non-content squares dependent on the presence of a similar square in different website pages. L-Extractor utilizes different square highlights and train a support vector (SV) based classifier to recognize an enlightening square versus a non-educational square. To begin with, the calculation

segment the site page into blocks dependent on heuristics. Second, the calculation groups each square as either a substance block or non-content square. It has the favorable position that both K-Feature Extractor and Content Extractor produce fantastic accuracy and review esteems and runtime effectiveness. It additionally decreases the multifaceted nature and expands the viability of the extraction cycle. It has the disservice that it will build the capacity necessity for lists and the productivity of the increase calculation are not improved.

Yinghui et al [3] proposed a strategy called Hierarchical example based bunching calculation. In light of utilizing thing sets to speak to designs in web exchanges, Greedy Hierarchical thing set based bunching (GHIC) has been introduced. From the outset, the arrangement of regular thing sets in the un-grouped information is acquired. From that point forward, another dataset (Binary thing sets dataset) was produced where the lines speak to the first exchanges and the sections speak to the presence or nonattendance of a successive thing set. This is spoken to as the new arrangement of exchanges. The issue was changed over into bunching these double vectors. At that point GHIC is introduced to take care of the bunching issue in the new arrangement of exchanges. It has the preferred position that a bunch of thing sets was permitted to depict a group rather than simply a bunch of things and it can clarify the cluster and the contrasts between groups. Contrast or likeness network was not considered here.

Jinbeom Kang et al [4] proposed another technique for site page division by perceiving label designs in the DOM tree structure of a page. These redundant HTML label designs are called key examples. Reiteration based page division (REPS) calculation is proposed to distinguish key examples in a page and to produce virtual hubs to effectively portion settled squares. REPS continues in four stages. Initial a page is spoken to by a DOM tree structure in the wake of eliminating less important labels, for example, <a>, <b>, <script>, and so forth from the HTML wellspring of the page. In the subsequent stage, REPS produces an arrangement from the DOM tree by utilizing the labels in the kid hubs of the root hub. The third stage is to locate the critical examples from the succession and perceive applicant blocks by coordinating the arrangement with the key examples. The last period of REPS is to produce blocks in a page by adjusting the DOM tree into an all the more profoundly various leveled structure by presenting virtual hubs.

Chia-Hui Chang [5] done a review of web data extraction frameworks. They saw a few focuses like to computerize the interpretation of information pages into organized information, a great deal of endeavors have been given in the region of information extraction (IE). IE creates organized information prepared for post handling, which is basic to numerous utilization of web mining and looking through instruments. The web IE measures online archives that are semi-organized and normally produce naturally by a worker side application program. Web IE typically applies AI and example mining methods to misuse the linguistic examples of the format based records. They discovered disservices like the extraction exactness is incredibly diminished in the event of absent or various request ascribes.

Wei Liu et al [6] proposed a strategy called Vision based methodology for profound web information extraction. It is basically founded on the visual highlights human clients can catch on the profound pages while additionally using some straightforward non-visual data, for example, information types and regular images to make the arrangement more powerful. It comprises of two fundamental segments, vision based information record extractor (ViDRE) and vision based information thing extractor (ViDIE). To begin with, given an example profound site page from a web information base, get its visual portrayal and change into a visual square tree. Second, remove information records from the Visual Block tree. Third, segment extricated information records into information things and adjust the information things of a similar semantic together. Fourth, produce visual coverings (a bunch of visual extraction rules) for the web information base dependent on example profound pages. ViDE can undoubtedly recognize the skewed information things because of their various textual styles or positions. Visual data of pages assists with executing web information extraction. Faults like either exactness or review isn't 100%. Likewise this measure demonstrates the level of web information bases the computerized arrangement neglects to accomplish amazing extraction.

Badr Hssina et al [7] proposed a strategy to separate required example by eliminating commotion that is available in the web report utilizing hand-made guidelines. Hand-created rules use string control capacities to separate data from HTML. Since the wellspring of data is a combination of picture, sound, introduction, and so forth, it is difficult to isolate out the instructive substance successfully and wisely. To begin with, interface with any site and get information from that site. At that point pick choices like concentrate joins, extricate picture, remove media, separate HTML outlines, separate substance.

R. Gunasundari [8] built up a strategy to extricate content from site pages, which depends on connections present in a site. In this technique, calculation makes a decision about the substance by a few boundaries in the

hubs. They are Link Text Density (LTD), Link Amount (LA), Link Amount Density (LAD) and Node Text Length (NTL). LTD and NTL are significant boundary for content area judgment and LA and LAD are pointers for exact substance judgment. The accompanying techniques are utilized for extricating the fundamental substance. To start with, normalize the website page labels. Second, pre-handling the site page labels. Third, making a decision about the area of the substance. Four, separating the substance. Five, changing the extraction results. This strategy can diminish amount of information transmission and unpredictability. Additionally it is reasonable for information assortment laborers and different experts. Idea recovery and the extension of semantic and equivalents are required for additional work.

As indicated by S.S. Bhamare [9] commotion on the site pages are not the piece of the primary substance and this insignificant data in site pages can truly influence web mining task. Two classes of commotion bunch are framed. They are worldwide clamor and neighborhood commotion. There are web cleaning procedures or strategies.

1. Page division physically or naturally portions a page into little squares zeroing in on rational subtopics.
2. Square coordinating recognizes sensibly similar squares in various pages.
3. Significance assessment gauges the significance of each square as indicated by various data or estimations.
4. Commotion assurance recognizes loud squares from non-boisterous squares dependent on the significance assessment of squares

Pralhad S Gamre et al [10] put together a bunch of archives into classes through grouping. Gathering of comparative records into bunches will assist the clients with finding the data without any problem. Items in a similar bunch ought to be comparative. Likewise, objects in a single bunch ought to be divergent from objects in other group. Mixture approach utilizes idea based mining model. In this model, it examinations terms on the sentence, record, corpus level and Hierarchical Agglomerative Clustering (HAC) to aggregate comparative reports in bunches and the archives are orchestrated in progressive structure to make simple access of web reports.

3. Deep Web Page Representation

The visual data of Web pages, which has been presented above, can be acquired through the programming interface gave by Web programs (i.e., IE). In this paper, we utilize the VIP calculation D. Cai et al [11] to change a profound Web page into a Visual Block tree and concentrate the visual data. Visual Block tree is really a division of a Web page. The root block speaks to the entire page, and each square in the tree relates to a rectangular district on the Web page Zehuan Cai [12] The leaf blocks are the squares that can't be divided further, and they speak to the base semantic units, for example, nonstop messages or pictures. Fig. 1a shows a well-known introduction structure of profound Web pages and Fig. 1b gives its comparing Visual Block tree. The specialized subtleties of building Visual Block trees can be found in D Jin Liu [13]. A genuine Visual Block tree of a profound Web page may contain hundreds even huge number of squares. Visual Block tree has three fascinating properties. To start with, block a contains block b if an is a progenitor of b. Second, a and b don't cover in the event that they don't fulfill property one. Third, the squares with a similar parent are orchestrated in the tree as indicated by the request for the relating hubs showing up on the page. These three properties are represented by the model in Fig. 1. The proper portrayals for inside squares and leaf blocks in our methodology are given underneath Mustafa Man [14] Jiachen Pu1 [15].

3.1. Visual Based Deep Web Data Extraction (VBDWDE) Algorithm

In spite of the fact that the conventional VBDWDE based calculations ordinarily portion the information districts effectively, it is as yet normal that the information territory of the page can be mistaken for different regions. There likewise exist different regions that contain data that we don't require, for instance, the route bar, promoting data, etc. We call them clamor in this article. This segment discloses how to eliminate the clamor when we direct the page division. Customary calculation, for example, VBDWDE has an excessive number of division rules. It is basic that the square division productivity is low. While the unpredictability of the page block continues to turn out to be more mind boggling, it's hard to direct enormous scope information extraction when confronting countless hubs of the Web page. To conquer the inadequacies of these calculations, CTVPS was proposed to utilize ordinary square labels, for example, ‹table› or ‹div›, for website pages division. Furthermore, this technique incredibly rearranges the heuristic principles. As the substance of a website page is completely enveloped by tag '‹table›' or tag '‹div›', using this labeling based technique enormously improve the proficiency of page division. Be that as it may, this strategy doesn't take out the futile mark data. We actually need alternate approaches to improve the extraction proficiency. By adding more principles to improve the productivity of the

DOM tree investigation and abatement the commotion, our technique, visual data based division can section the pages proficiently. The VIBS calculation depends on the VIPS calculation. Some improvement proposed by CTVPS is likewise included. The CTVPS calculation has added some new section rules, for example, based on ‹table› or ‹div› labels to accelerate the productivity of the page block crossing, instead of navigating every hub of the DOM tree. Along these lines, based on CTVPS calculation, we consolidate the properties of DOM hubs with the related visual data that is acquired by our site page picture preparing model. We additionally add heuristic guidelines for eliminating brightened labels, for example, ‹font›, ‹i›, ‹strong›, etc. These labels are utilized to feature key data, for instance, the cost of merchandise, titles, etc. At the point when we split the page to assemble a visual square tree, area data of all adjusted labels is recorded and saved. In the wake of finding the information area, we eliminate the adjusted hub and move its youngster hub up as its correct kin. Expulsion of the adjusted mark right now will impact the area of the information region, so we just record the relating area of the altered name data. Subsequent to adding extra heuristic standards, the strategy can build a moderately silent square tree. The propose VBDWDE calculation as appeared in the follows.

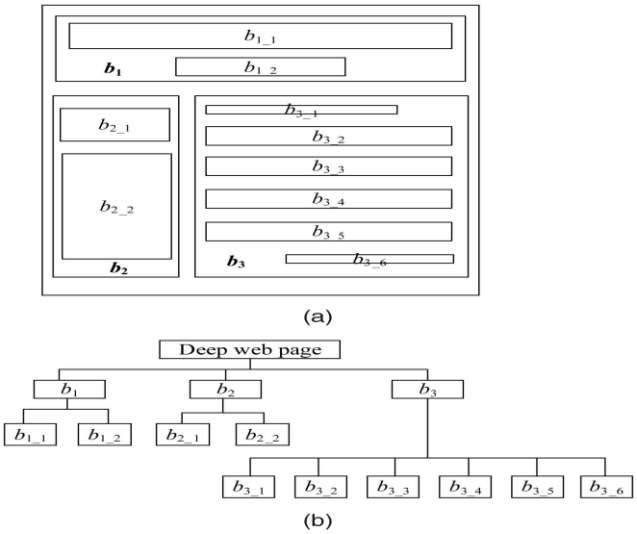


Figure 1. (a) The presentation structure and (b) Its Visual Block tree

Algorithm 1: VBDWDE Algorithm

```
1: Input: the whole web page, whole web page image
2: Begin:
      Load(heuristic rule library)
      GenerateDomTree(pNode, nLevel)
      {
  If(isDividable(pNode,nLevel)==(<table>or<div>))
      {
              For each child of pNode
              {
                      If(find(tagNode)){
                      getNodePosition(tagNode);
                      map.put(tagNodeName,position);
                }
GenerateDomTree(child,nLevel);
       }
    }
}
Else
{
 Put the sub-tree(pNode) into the pool
as a visual block;
}
Load(data location region model)
RegionResult = Model(whole web page image)
 VisionResult = ExtractDomTree(RegionResult)
```

Compare(visual block, VisionResult)
{
Return HighConfidenceBlock
}
End
3: Output: Classification Result

As appeared in the calculation, initially we load a heuristic principle library. At that point we accept the whole page as contribution of the calculation, by following similar strides as CTVPS, we at that point develop the ‹table› or ‹div› label tree, separate the page block, recognize the section, build the label tree, extricate the page block, develop the semantic impede and get the visual square tree.

4. Result and discussion

The Performance of the Visual Base Deep Web Data Extraction is evaluated by using web databases. The experimental assessments are carried out to check the web pages extracted from deep web are discovering expected amount of information. These web databases are selected from several domains and different set of users submit their queries to access the appropriate information from the web databases. Three set of user queries are offered and collect five deep Web pages holding three data records at any time, for each web database. The proposed Vision based Deep Web Data approach is compared with the existing ViDE, FiVA Tech and VBEC to access the visual information from the web databases offering better speed and high precision.

4.1. Result Analysis

Results are evaluated and analysed under below mentioned criteria and presented in graphs given below. Each criterion is evaluated separately and presented.

Table 1. Precision Measurement

No of Entries in Web Database	**Precision**			
	ViDE	**FiVA Tech**	**VBEC**	**VBDWDE**
100	68	75	83	90
200	72	77	77	89
300	75	80	88	91
400	80	85	89	92
500	82	89	91	94
600	86	90	92	96

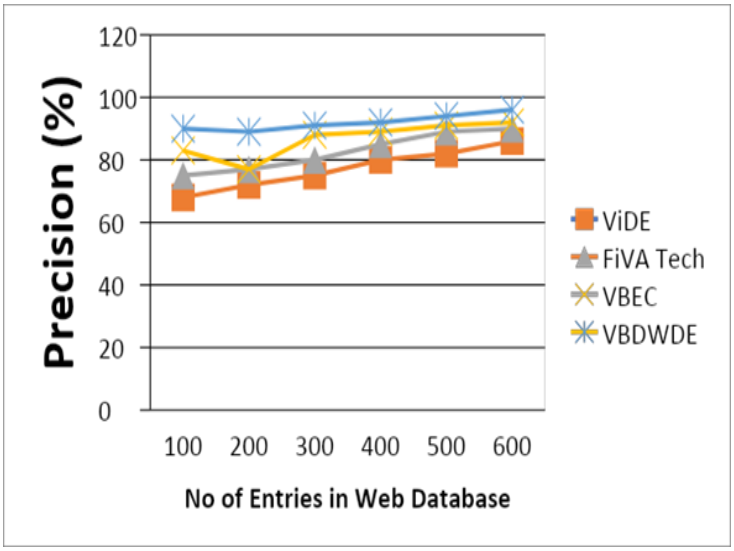


Table 2. Speed Measurement

No of Entries in Web Database	Time Consumption(ms)			
	ViDE	FiVA Tech	VBEC	VBDWDE
100	22	20	15	10
100	29	26	19	13
200	35	31	26	20
300	40	36	30	22
400	48	43	34	29
500	52	50	40	31

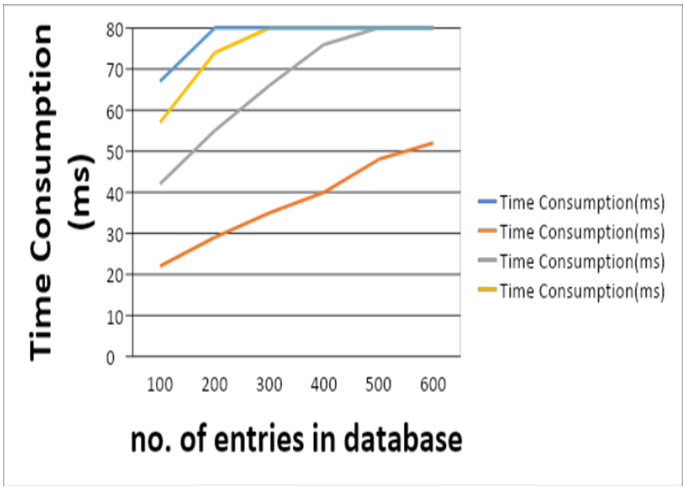
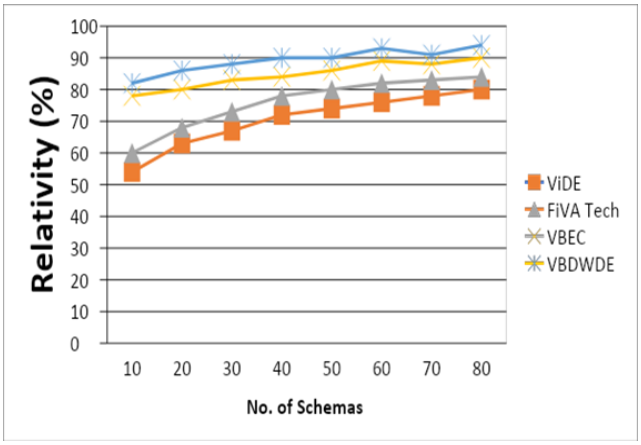


Table 3. Relativity Measurement

No of Schemas	Relativity(%)			
	ViDE	FiVA Tech	VBEC	VBDWDE
10	54	60	78	82
20	63	68	80	86
30	67	73	83	88
40	72	78	84	90
50	74	80	86	90
60	76	82	89	93
70	78	83	88	91
80	80	84	90	94



Finally, proposed solution VBDWDE achieved the data extraction on users request. The performance criterion such as precision, speedy access and greater relativity are fulfilled.

References

1. A.F.R Rahman, H.Alam and R.Hartono, "Content extraction from HTML documents", International workshop on Web document Analysis, pp.7-10, 2001.
2. Sandip, prasenjit, Nirmal Pal and C.Lee Giles, "Automatic Identification of Informative Sections of web pages", IEEE Transactions on knowledge and data Engineering, Vol 7, No 9, 2005.
3. Yinghui Yang and Balaji Padmanabhan "A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions", IEEE Transactions on knowledge and data Engineering, Vol 7, No 9, 2005.
4. Jinbeom Kang, Jaeyoung Yang, Nonmember and Joongmin Choi, "Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices", IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010.
5. Chia-Hui Chang, Moheb Ramzy Girgis, "A Survey of Web Information Extraction Systems", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 10, October 20006.
6. Wei Liu, Xiaofeng Meng and Weiyi Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2010.
7. Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali, Belaid Bouikhalene "An implementation of web content extraction using mining techniques", Journal of Theoretical and Applied Information Technology 31st December 2013. Vol. 58 No.3, ISSN: 1992-8645.
8. R.Gunasundari, "A study of content extraction from web pages using links"International Journal of Data Mining & knowledge management process, Vol.2, No.3, May2012.
9. S.S. Bhamare, Dr.B.V., "Survey on Web Page Noise Cleaning for Web Mining", International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013.
10. Pralhad S. Gamare, G. A. Patil, "Web document clustering using hybrid approach in data mining" International Journal of Advent Technology, Vol.3, No.7, July 2015.
11. D. Cai, S. Yu, J. Wen, and W. Ma, "Extracting Content Structure for Web Pages Based on Visual Representation," Proc. Asia Pacific Web Conf. (APWeb), pp. 406-417, 2003.
12. Zehuan Cai, Jin Liu, Lamei Xu, Chunyong Yin, Jin Wang "A Vision Recognition Based Method for Web Data Extraction" Advanced Science and Technology Letters Vol.143 (AST 2017), pp.193-198 http://dx.doi.org/10.14257/astl.2017.143.40.
13. Jin Liu ,Li Lin , Zehuan Cai , Jin Wang, Hye-jin Kim, Deep web data extraction based on visual information processing, Journal of Ambient Intelligence and Humanized Computing (2017).
14. Mustafa Man, Ily amalina ahmad sabri, Wan Aezwani Bt Wan Abu Bakar Web Data Extraction Approach for Deep Web using WEIDJ.
15. Jiachen Pu1 , Jin Liu1(&) , and Jin Wang, A Vision-Based Approach for Deep Web Form Extraction, International Conference on Multimedia and Ubiquitous Engineering FutureTech 2017, MUE 2017: Advanced Multimedia and Ubiquitous Engineering pp 696-702.

Authors

Shilpa A. Deshmukh (shilpad@sies.edu.in) graduated from Shivaji Science College , Amravati University of Amravati, India. She did her M.C.A. (Master of Computer Application), from DCPH of HVPM, Amravati University. She is now an Assistant Professor of M.C.A. Department, SIESCOMS, Navi Mumbai. Her research activities span from Data mining, Web Mining, Deep Web Mining.

Dr. Pravin P.Karde was born in Amravati, Maharstra in 1975. He received the Post Graduate Degree (M.E.) in Computer Sceince & Engineering from S.G.B. Amravati University, Amravati in the year 2006 & Completed the Ph.D degree in Computer Science & Engineering. Currently he is working as an Assistant Professor in Information Technology Department at Government Polytechnic, Amravati, ,India. His interest is in

Selection &Maintenance of Materialized View.

Dr. V.M.Thakare was born in Wani, Maharashtra in 1962. He had worked as Assistant Professor for 10 Years at Professor Ram Meghe Institute of Technology & Research, Badnera and P.G.Department of Computer Science, S.G.B. Amravati University, Amravati. Currently he is working as Professor & Head in Computer Science from last 9 years, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, SGB Amravati University, Amravati.. He has published 86 papers in various National & International Conferences & 20 papers in various International journals. He is working on various bodies of Universities as a chairman & members. He has guided around 300 more students at M.E / MTech, MCA M.S & Mphil level. He is a research guide for Ph.D. at S.G.B. Amravati University, Amravati. His interests of research is in Computer Architecture, Artificial Intelligence, Robotics, Database and Data warehousing & mining.