IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Interactive Data Extraction Algorithm to Extract Data from The Pdf Document, Helpful in Generating Water Quality Data of The Kanhan River

D. A. Lingote[1*], Girish S. Katkar[2]

[1]CSIR-National Environmental Engineering Research Institute, Nehru Marg, Nagpur, India
[2]Department of Computer Science and Application, Art, Commerce & Science College, Koradi, Nagpur, India

[*]*Corresponding Author: da_lingote@neeri.res.in Tel.: +91-7773952770*

DOI: https://doi.org/10.26438/ijcse/v7i3.550556 | Available online at: www.ijcseonline.org

Accepted: 10/Mar/2019, Published: 31/Mar/2019

Abstract— Now a day's internet is very popular and widely used for information generation and broadcasting. If current trend is observed, then most of the organization/labs/institute uses "PDF" (Portable Document Format) document to release their official/research report. PDF document has many benefits, hence popularly used for publishing information on the web. if this widely published information extracted and re-processed then this information can be useful inputs for many research and development projects. In this research paper we introduced information extraction algorithm, which extracts information from the pdf document using free libraries. To be specific, we have targeted PDF documents comprising Kanhan River water quality data, which is freely published over the internet. To present this information beautifully, extracted information is geo-mapped and re-published in the public domain which helps in observing and validating Kanhan River water quality data at different geographical locations.

Keywords— PDF Extraction, data generation, Extraction, Kanhan River, information system

I. INTRODUCTION

Pdf reader is available free for the download. PDF files are printable and viewed on any operating systems like: Unix, android, windows and Mac, hence can also be easily readable on mobiles even. PDF document can be secured using password and disseminated restricting copy-paste. PDF document can be made searchable defining metadata. Fonts of the document remains embedded in the document, hence ensures its readability at the second end. Pdf document uses a compressed format, hence if images are used in the document, then in such case file size remains relatively low. Having such benefits, PDF file is most popularly and widely used file format for generating and disseminating information.

Research organizations periodically conducts sampling drive and generate water quality data for the Kanhan River. Maharashtra Pollution Control Board (MPCB, Mumbai) also conducts water quality study drive periodically at different geographical locations to assess river water quality. Indeed, MPCB has set-up several Water Quality Monitoring Stations (WQMS) on the rivers, among these around 3-4 WQMS have been set-up on the Kanhan river. Assessed water quality either will be published on their websites or published by means of research reports and such data is targeted for the extraction to generate central water quality information storage for the Kanhan River.

II. RELATED WORK

Dr. G. K. Khadse and his research team has carried out kanhan river study and generate water quality data for the Kanhan River at different geographical locations for the year 2006 and 2007 [1]. Authors of this research paper are involved in generating database of the Kanhan River and explored different methodologies for the data generation. Team has targeted different information sources available in public domain for the extraction [2] & [3] and utilized in information generation. However, data extraction from the pdf document for the Kanhan River information generation is unique of its kind.

III. METHODOLOGY

The Kanhan River originates from the Satpura hills situated in Chhindwara districts of Madhya Pradesh, river reaches to Nagpur district of Maharashtra and confluences the Wainganga River. Around 300Kms of Kanhan River stretch from Amla Dam, Madhya Pradesh to Gosekhurd Dam, Maharashtra, India is considered along with water sources confluences the river to generate database for the kanhan River. In view to significantly notify exact sampling points

on the river path, Study Points (SPs) are imposed on the river at 100-meter distance. Figure-1 shows the Kanhan River path considered for the information generation. Different data generation methodologies like information extraction, generation and estimation are used to generate information of the Kanhan River. Whereas, this paper focuses on the algorithm developed for extracting data from the PDF document.

River engineering requires numerical data; hence only structured numerical data is considered for the extraction. However, developed algorithm can extract all sorts of text and numbers (text, Semi-numeric data and tabular-numeric data) from the input stream. Developed algorithm ignores images, charts, block-diagram etc. while extraction.

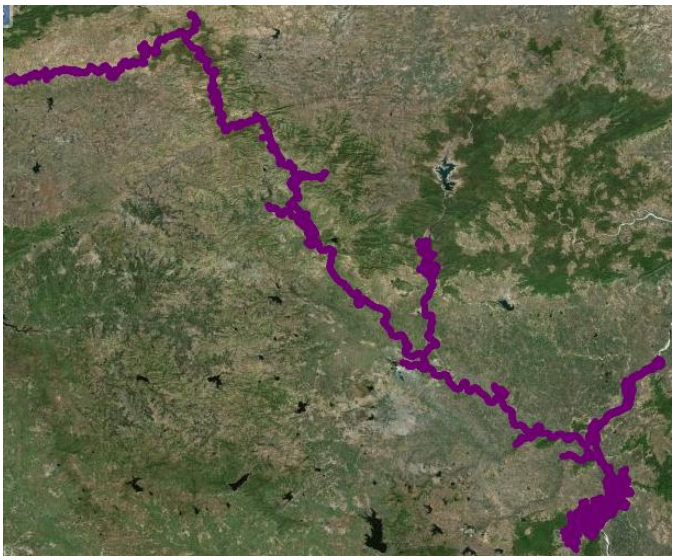

Figure 1. Complete path of the Kanhan River considered for the information generation

"org.apache.pdfbox.*" library[4] is popular for reading pdf docment, which helps in reading data from the pdf document. Freely downloadable library provides moderate features to read PDF document, hence used here for the extraction. Java technologies are used for the development of algorithm and PostgreSQL database server is used as a backend. Referral latitude and longitude of Google Earth are used for the global positioning of the information on the map. Workspace of PostgreSQL database server is mapped with Geosever 2.11 and information layers are published using Tomcat web server. Developed algorithm uses following procedures for the data extraction.

In order to explain the algorithm, here we have extracted data from two source PDF documents which are Water Quality Index (WQI, published by MPCB) document published in public domain on the website of MPCB, Mumbai and data generated by CSIR-NEERI. Figure 2 shows tree structure used for the information extraction.

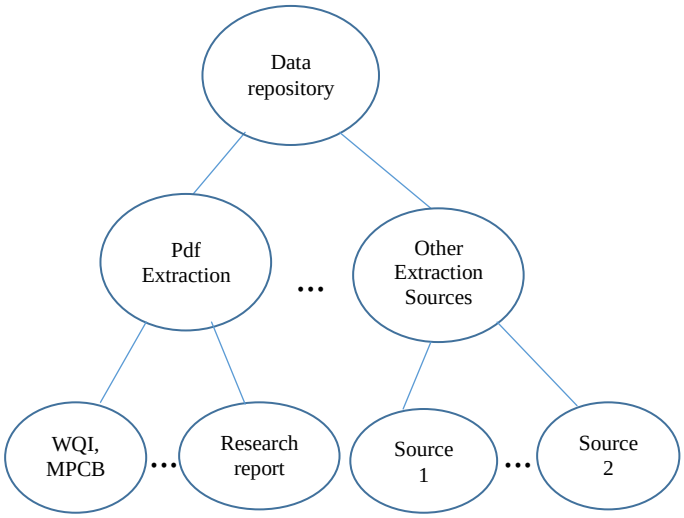

Figure 2. Tree structure used for data generation.

Algorithm:

As shown in Figure 3, algorithm uploads PDF file to be extracted into the system folder. Figure 4 shows the list files uploaded into system folder and against each file (extraction column) extraction icon is given, which can be used to apply the developed extraction algorithm. Along with, algorithm also provide feature to edit file details or delete file from the system folder if not required. Figure 4 shows the other extraction options which appears when user clicks the extraction icon. As shown in figure 5, algorithm facilitates two methods for extracting pdf documents "Extract" and "Table Extract" respectively.

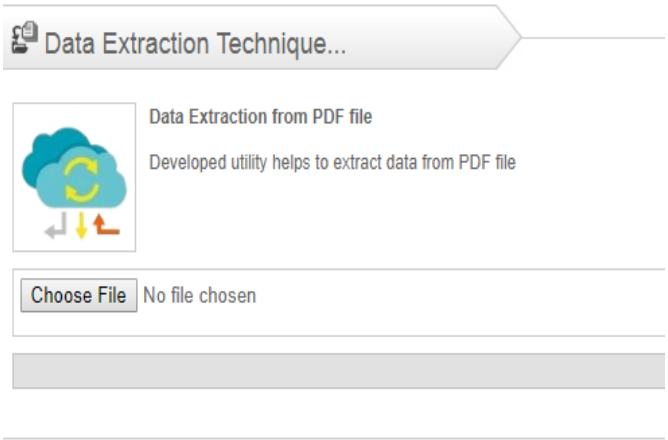

Figure 3. Upload file to be extracted into the system folder.

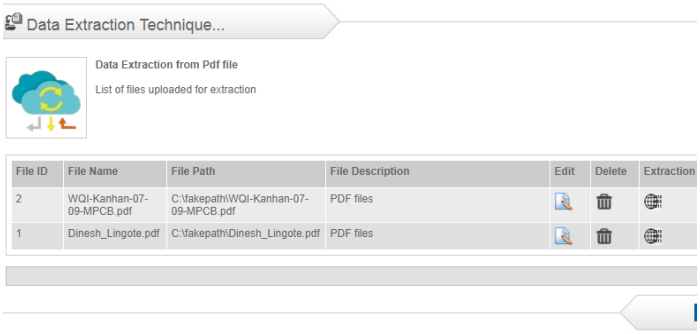

Figure 4. shows the list files uploaded in system folder

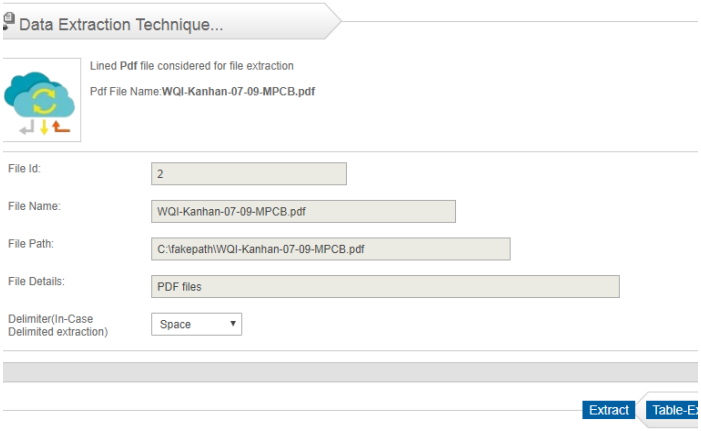

Figure 5. Extraction options

1. Extract

This option extracts input document line by line till the end of line (till new-line character) and write the line/paragraph into the Excel document. Algorithm treats every encountered input line as a paragraph and continues to read the input stream line by line till end of the file and generate Excel document accordingly. This option is suitable for the input file when it is purely in text/paragraph format.

2. Table Extract

This extraction option uses splitters like: space (" "), Semicolon (;), colon (:) and (,) are used to separate input line into different columns of the Excel document. Wherein space splitter is best suitable option for tabular-numeric data extraction, Semicolon (;) option is suitable when input data is separated with semicolon, option colon (:) is suitable when input data is separated with colon and lastly splitter option comma (,) is used for the input file in which Comma Separated value(CSV) are available

Here we have considered two documents for the extraction and accordingly resulted documents are depicted after applying appropriate extraction algorithm. Figure 6 shows the first Input Water Quality Index(WQI) document and figure 7 shows the second input Research-report document considered for the extraction. Subsequently, figure 8 shows the list extracted Microsoft Excel files, which got generated after applying the appropriate extraction algorithm.

Chapter 4

Evaluation of River Water Quality

4.1 Surface Water Monitoring Network

For the years 2007-09, monitoring of water quality was carried out under various programmes namely NWMP, SWMP and Hydrology project. The present chapter covers all the major and minor rivers of Maharashtra considered for water quality monitoring. **Table 4.1** gives the list of the rivers and their number of stations considered for water quality analysis.

Table 4.1: Major and Minor Rivers Covering Surface Water Monitoring Network

Sr.	River Name	No. of Stations	Sr.	River Name	No. of Stations
1	Amba	2	31	Muchkundi	2
2	Amravati	1	32	Mula-Mutha	8
3	Bhatsa	3	33	Nira	5
4	Bhima	8	34	Penganga	2
5	Bindusara	1	35	Panchganga	4
6	Bori	1	36	Panzara	1
7	Burai	1	37	Patalganga	8
8	Chandrabhaga	2	38	Pawana	7
9	Daman Ganga	3	39	Pedhi	1
10	Darna	5	40	Pelhar	1
11	Deoghar	1	41	Purna	3
12	Ghod	1	42	Rangavali	1
13	Girna	2	43	Savitri	7
14	Godavari	51	44	Sina	1
15	Gomai	1	45	Shastri	1
16	Hiwara	1	46	Surya	3
17	Indravani	3	47	Tansa	1
18	Kajvi	2	48	Tapi	21
19	Kalu	1	49	Titur	1
20	Kan	1	50	Ulhas	6
21	Kanhan	3	51	Urmodi	1
22	Kodavali	1	52	Vaitarna	5
23	Kolar	1	53	Vashishti	4
24	Koyna	1	54	Vel	1
25	Krishna	29	55	Venna	3
26	Kundalika	5	56	Waghur	1
27	Manjra	1	57	Wardha	7

Figure 6. Water Quality Index document (First input document)

pH	Cond. (µS/cm)	Ca (meq/L)	Mg -	Na -	K -	Cl -	SO4 -
7.1	326	1.3	1.7	1.3	0.2	0.54	0.17
7.2	336	1.4	1.7	1.3	0.3	0.68	0.21
8.0	370	1.2	1.9	1.2	0.3	0.62	0.25
7.7	431	1.5	2.0	1.5	0.4	0.76	0.33
8.2	383	1.5	1.6	1.7	0.5	0.62	0.38
8.2	432	1.5	1.6	1.8	0.5	0.65	0.42
8.2	394	1.6	1.6	1.9	0.4	0.79	0.35
8.3	412	1.5	1.7	2.0	0.5	0.82	0.38
8.1	400	2.0	1.2	2.0	0.3	1.01	0.27
8.2	413	2.1	1.1	2.2	0.4	1.04	0.33
8.2	406	1.7	1.7	2.2	0.4	1.21	0.33
8.3	408	1.7	1.6	2.3	0.5	1.30	0.39
8.3	604	2.1	1.6	3.9	0.3	1.49	0.25
8.3	632	2.3	1.8	4.0	0.3	1.72	0.25
7.7	262	1.3	1.7	0.9	0.3	0.85	0.23
7.8	234	1.3	1.7	1.0	0.2	0.96	0.19
7.6	250	1.6	1.4	1.0	0.3	0.79	0.25
7.6	227	1.6	1.2	1.2	0.4	0.90	0.29
8.1	279	1.5	1.4	1.6	0.4	0.70	0.31
8.2	318	1.5	1.6	1.7	0.4	0.68	0.33
8.5	970	2.6	2.2	5.3	0.6	2.42	0.46
8.6	888	2.5	2.3	5.5	0.6	2.51	0.48
8.6	848	3.0	1.7	6.7	0.5	2.48	0.42
8.7	852	3.1	1.9	7.3	0.5	2.73	0.40
8.7	823	3.0	1.6	7.7	0.4	2.82	0.29
8.5	804	3.1	1.6	8.0	0.4	2.87	0.33
7.8	349	2.1	1.1	2.3	0.4	1.35	0.29
7.4	412	2.0	1.2	2.7	0.4	1.58	0.31

Figure 7 Research report document (Second input document)

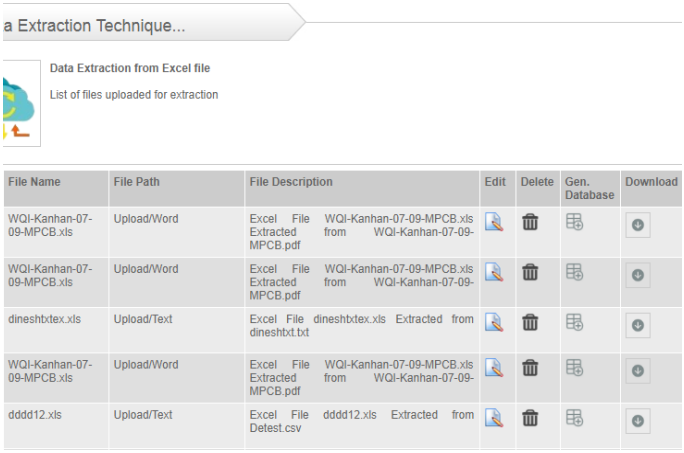

Figure 8. List of Excel files generated by the algorithm applying selected extraction algorithm

IV. RESULTS AND DISCUSSION

As explained, we have applied both the extraction algorithm on two different pdf documents. Figure 9 shows the result of extraction when paragraph extraction technique applied on WQI document. Figure 10 shows the result of extraction when Table extraction technique applied on WQI document. Similarly, Figure 11 shows the result of extraction when paragraph extraction technique is applied on Research report. Lastly, figure 12 shows the result of extraction when Table extraction technique is applied on Research report.

A	B	C	D	E	F	G	H
4-1							
Chapter 4							
Evaluation of River Water Quality							
4.1 Surface Water Monitoring Network							
For the years 2007-09, monitoring of water quality was carried out under various programmes namely NWMP, SWMP and Hydrology project. The present chapter covers all the major and minor rivers of Maharashtra considered for water quality monitoring. Table 4.1 gives the list of the rivers and their number of stations considered for water quality analysis.							
Table 4.1: Major and Minor Rivers Covering Surface Water Monitoring Network							
Sr. River Name No. of Stations Sr. River Name No. of Stations							
1 Amba 2 31 Muchkundi 2							
2 Amravati 1 32 Mula-Mutha 8							
3 Bhatsa 3 33 Nira 5							
4 Bhima 8 34 Penganga 2							
5 Bindusara 1 35 Panchganga 4							
6 Bori 1 36 Panzara 1							
7 Burai 1 37 Patalganga 8							
8 Chandrabhaga 2 38 Pawana 7							
9 Daman Ganga 3 39 Pedhi 1							
10 Darna 5 40 Pelhar 1							
11 Deoghar 1 41 Purna 3							
12 Ghod 1 42 Rangavali 1							

Default Sheet

Figure 9. Extraction result when applied paragraph extraction technique – WQI

	A	B	C	D	E	F	G
1	4-1						
2	Chapter	4					
3							
4	Evaluation	of	River	Water	Quality		
5	4.1	Surface	Water	Monitoring	Network		
6							
7	For	the	years	2007-09,	monitoring	of	water
8	namely	NWMP,	SWMP	and	Hydrology	project.	The
9	minor	rivers	of	Maharasht	considerec	for	water
10	the	rivers	and	their	number	of	stations
11							
12							
13							
14	Sr.	River	Name	No.	of	Stations	Sr.
15	1	Amba		2	31	Muchkund	2
16	2	Amravati	1	32	Mula-Muth	8	
17	3	Bhatsa	3	33	Nira	5	
18	4	Bhima	8	34	Penganga	2	
19	5	Bindusara	1	35	Panchgan	4	
20	6	Bori	1	36	Panzara	1	
21	7	Burai	1	37	Patalgang	8	
22	8	Chandrabh	2	38	Pawana	7	
23	9	Daman	Ganga	3	39	Pedhi	
24	10	Darna	5	40	Pelhar	1	
25	11	Deoghar	1	41	Purna	3	
26	12	Ghod	1	42	Rangavali	1	

Figure 10. Extraction result when applied Table Extraction technique -WQI

	A	B	C	D	E	F
22	8 Pench On Chhindwara Chaurai Road, <12 km from Chaurai 79°0					
23	9 Pench Near Navegaon Khairi village 78°56'55" 21°32'10" 336					
24	10 Pench Near Bina village before confluence with Kanhan river 79°					
25	11 Nag Near Bharatwada village in the east of Nagpur city 79°08'12					
26	12 Nag On Kuhi Badoda Road 79°22'01" 21°02'05" 260					
27	13 Nag Near Panmara village before confluence with Kanhan river 7					
28	14 Wainganga Near Ambhora village after confluence with Kanhan					
29	S.N. pH					
30	Turb.					
31	(NTU)					
32	Cond.					
33	µS/cm)					
34	TDS					
35	T.					
36	Alk.					
37	T.					
38	Hard. Ca Mg Na K Cl SO4 NO3 PO4 DO COD TC FC					
39	(mg/L) - - - - - - - - - - - - (CFU/100ml)					
40	1 7.1 0.8 326 196 188 156 26 21 29 4 19 8 4 0.2 8.5 7 154 48					
41	- 7.2 1.7 336 202 191 158 28 21 30 5 24 10 5 0.1 7.6 9 106 28					
42	2 8.0 1.3 370 222 204 164 24 23 28 4 22 12 3 0.3 7.2 10 80 15					
43	- 7.7 1.0 431 259 212 168 29 24 34 7 27 16 4 0.4 7.0 13 112 28					
44	3 8.2 1.4 383 230 210 157 30 20 38 9 22 18 6 0.4 7.3 12 84 21					
45	- 8.2 1.6 432 259 216 156 30 20 41 10 23 20 5 0.2 7.2 16 36 15					
46	4 8.2 1.6 394 236 212 160 31 20 44 7 28 17 6 0.1 6.9 17 47 18					

Figure 11. Extraction result when applied paragraph extraction technique – Research Report

A	B	C	D	E	F	G	H	I	J	K	L
5	Kanhan	Near	Mathni-Ma	on	NH-6	79°23'17"	21°08'45"	251			
6	Kanhan	Near	Panmara	village	before	confluence	with	Nag	river	79°28'01"	21°05'36"
7	Kanhan	Near	Panmara	village	after	confluence	with	Nag	river	79°28'03"	21°05'34"
8	Pench	On	Chhindwar	Chaurai	Road,	<12	km	from	Chaurai	79°09'53"	22°02'41"
9	Pench	Near	Navegaon	Khairi	village	78°56'55"	21°32'10"	336			
10	Pench	Near	Bina	village	before	confluence	with	Kanhan	river	79°10'15"	21°16'26"
11	Nag	Near	Bharatwad	village	in	the	east	of	Nagpur	city	79°08'12"
12	Nag	On	Kuhi	Badoda	Road	79°22'01"	21°02'05"	260			
13	Nag	Near	Panmara	village	before	confluence	with	Kanhan	river	79°28'01"	21°05'35"
14	Waingang	Near	Ambhora	village	after	confluence	with	Kanhan	river	79°24'58"	20°53'36"
S.N.	pH										
Turb.											
(NTU)											
Cond.											
	µS/cm)										
TDS											
T.											
Alk.											
T.											
Hard.	Ca	Mg	Na	K	Cl	SO4	NO3	PO4	DO	COD	TC
(mg/L)	-	-	-	-	-	-	-	-	-	-	-
1	7.1	0.8	326	196	188	156	26	21	29	4	19
-	7.2	1.7	336	202	191	158	28	21	30	5	24
2	8.0	1.3	370	222	204	164	24	23	28	4	22
-	7.7	1.0	431	259	212	168	29	24	34	7	27
3	8.2	1.4	383	230	210	157	30	20	38	9	22

Figure 12. Extraction result when applied Table extraction technique –Research Report

Developed algorithm generate a system file of the extracted data, which uses geo-references (latitude and longitude) given in the input document and accordingly geo-maps the input data. If Research-report is considered, then extracted data belong to seven different geographical locations on the Kanhan River path namely: Sausar, Chhindwara road (closed to Berdi, First), near Khapa closed to Parseoni road (Second), near Bina before confluence of the Pench River(Third), near Bina after confluence of the Pench River(Fourth), Mathni-mauda (Fifth), Near Panmara village before confluence of the Nag river(Sixth) and Near Panmara village after confluence of the Nag river(Seven). Apparently, figure 13 shows the geo-mapped extracted data near Sausar, Chhindwara road (closed to Berdi) for the year-2006. Figure 14 shows extracted geo-mapped data near Khapa, Parseoni (closed to Tembhurdoh) for the year-2007. Figure 15 shows extracted geo-mapped data near Bina before and after confluence of the Pench River for year 2006. Figure 16 shows extracted geo-mapped data near Mathani for the year 2007. Figure 17 shows extracted geo-mapped data near Panmara before and after confluence of the Nag River for the year 2007. Figure 18 shows extracted geo-mapped data near Panmara before and after confluence of the Nag River for the year 2006.

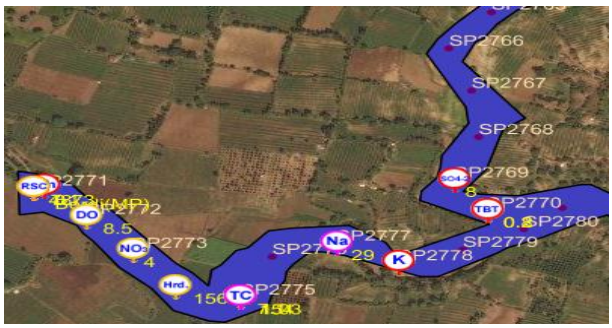

Figure-13. extracted & geo-mapped data near Sausar, Chhindwara road (closed to Berdi), Year-2006

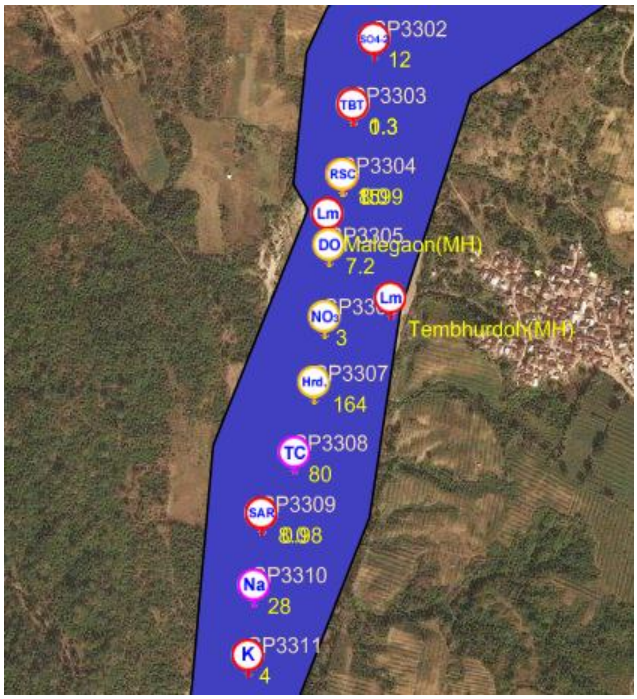

Figure 14. shows extracted geo-mapped data near Khapa, Parseoni (closed to Tembhurdoh) year-2007

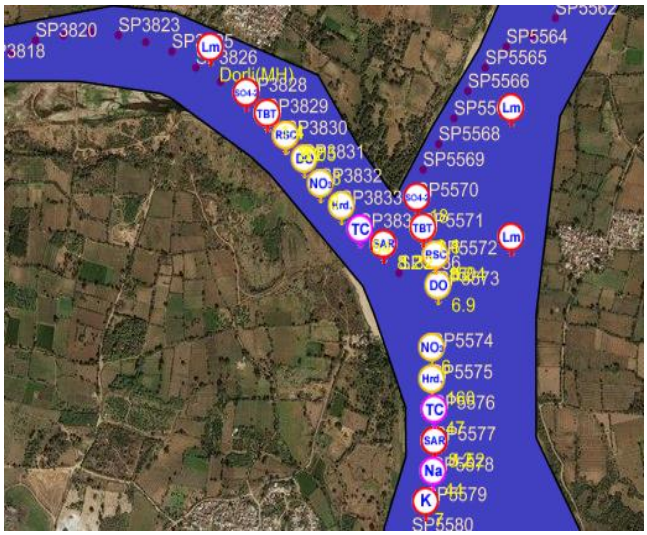

Figure 15. shows extracted geo-mapped data near Bina before and after confluence of the Pench River Year 2006

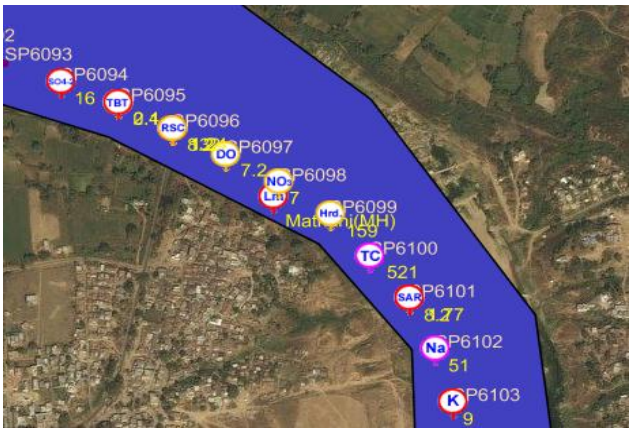

Figure 16. shows extracted geo-mapped data near Mathani for the year 2007

Figure 17 shows extracted geo-mapped data near Panmara before and after confluence of the Nag River year 2007

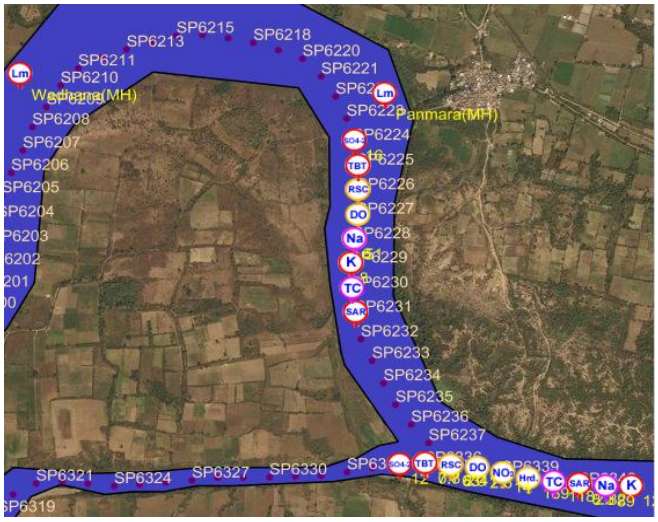

Figure 18. shows extracted geo-mapped data near Panmara before and after confluence of the Nag River year 2006

Discussion:

"org.apache.pdfbox.*" library is used to read pdf document, library has no specific feature to read the tables comprised in the document. Therefore, to solve this issue space delimiter is used for extracting data. This results in misappropriation, if cell of the table contains data with space (Figure 10). Introduced algorithm treats the space input character appears before the data-item (or in between the data items) in a cell of the table as separate items and split such data-item in two separate columns. Developed algorithm is best suitable to numeric-tabular data only and such data only targeted for the extraction as it is main input for the river study. To resolve such issues auto data correction techniques of excel is used, which helps in minor tuning of the document. If required, minor editing may also be carried out to finalize the extracted document, thus involves the manual intervention in data extraction. Significant development in the libraries (which help in reading pdf document) is going-on, which may help in reading tables from the pdf document and ultimately will be helpful in reading pdf document automatically.

If compared, the volume of data available in the open domain and the volume of data required from these large sources, then it realizes that volume of required data is very low which required to be extracted from different available sources. Therefore, although extraction techniques ease the data generation burden, beside it ultimately burdens the data processing team. Having such awareness, research team focuses to develop auto data processing techniques which can even ease the burden of data processing team.

V. Conclusion and Future Scope

If considered aim (extract numeric-tabular data) and the results generated by the extraction algorithm, then it can be significantly notified that introduced algorithm gives above 90 percent accuracy while extracting numeric-tabular data from the input stream. As discussed, introduced concept has wide scope in generating information. It's a general phenomenon that every user generates a data as per his goals and mission. Whereas other user (research team) may find it incomplete according to his work. The extraction collects substantial amount of unwanted data along with useful data, hence it ultimately gives major scope for generating automated technique to auto process extracted along with auto extraction.

Acknowledgment

I wish to express my sincere gratitude to Dr Rakesh Kumar, Director CSIR-NEERI for providing me an opportunity to complete Ph.D. work. I sincerely thank to Dr Girish S. Katkar, Professor and Head, Department of Computer Science and Application, Art, Commerce & Science College, Koradi, Nagpur for his guidance and support to carry out this research work.

REFERENCES

[1] Dr. G. K. Khadse, P. M. Patni, P.S. Kelkar, S. Devotta, "Qualitative evaluation of Kanhan River and its tributaries flowing over central Indian plateau", Environ Monit Assess. 2008 Dec; 147 (1-3):83-92. Epub 2007 Dec 22.

[2] Margaret H. Dunham, "Data Mining Introductory & Advanced Topics", Pearson Education

[3] Dinesh A. Lingote1*, Girish S. Katkar2, Ritesh Vijay 3, R. B. Biniwale4, "Responsive Information generation system for Kanhan River, an effective information system for river modeling", International Journal of Computer Science and Engineering (IJCSE, E-ISSN: 2347-2693), Vol.-6, Issue-12, Dec 2018

[4] Library org.apache.pdfbox.* is attributed as it is used for reading PDF document.

[5] Mehrdad Jalali, Norwati Mustapha et al," A Recommender System Approach for Classifying User Navigation Patterns Using Longest Common Subsequence Algorithm", American Journal of Scientific Research ISSN 1450-223X Issue 4 (2009), pp 17-27

[6] K. A. Smith and A. Ng, Web page clustering using a self-organizing map of user navigation patterns, Decision Support Syst. 35(2) (2003) 245–256

[7] Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles "A Comparison of Dimensionality Reduction Techniques for Web Structure Mining", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, P.116-119 ,2007

[8] Poonam Devi, "Attacks on Cloud Data: A Big Security Issue", International Journal of Scientific Research in Network Security and Communication, Volume-6, Issue-2, April 2018

[9] P.V. Nikam, D.S. Deshpande, "Different Approaches for Frequent Itemset Mining", International Journal of Scientific Research in computer science and Engineering, Vol.6, Issue.2, pp. 10-14, April (2018)

Authors Profile

Mr. Dinesh A. Lingote pursed B.Sc. (Computer Science) from Amravati Univeristiy, pursed M.C.A. from Amravati university and currently pursuing Ph.D in RTMNU, Nagpur, Under the Guidance of Dr. Girish Katkar, Department of Computer Science and Application, Art, Commerce & Science College, Koradi, Nagpur. He is currently working as Sr. Scientist in CSIR National Environmental Engineering Research Institute, Nehru Marg, Nagpur 440020. CSIR-NEERI is a Constituent laboratory of CSIR (Council of Scientific and Industrial Research, Anusandhan Bhawan, 2 Rafi Marg, New Delhi, Delhi 110001) The Renowned Institution is emerged as significant solution provider in the area of environmental science and engineering for the sustainable development. CSIR-NEERI have highly qualified and experienced scientist and engineers, who have developed several technologies and published quality research papers in environmental science and engineering area. He is not only involved in research and development, but also administers IT infrastructure of the Institution. He is currently working on auto data generation and extraction techniques to generate large data for environmental science and engineering.

Dr Girish S. Katkar, pursed M.Sc., PGDCS, MCM, He has pursed Ph.D. in computer science from RTMNU, Nagpur. He is currently working as Professor and Head of Department of computer science and application, art, Commerce & Science College, Koradi, Nagpur. He has guided several Ph.D. scholars, and published several quality research papers. He is currently working on data mining.

