REVIEW OF RESEARCH
IMPACT FACTOR : 5.7631(UIF) UGC APPROVED JOURNAL NO. 48514 ISSN: 2249-894X

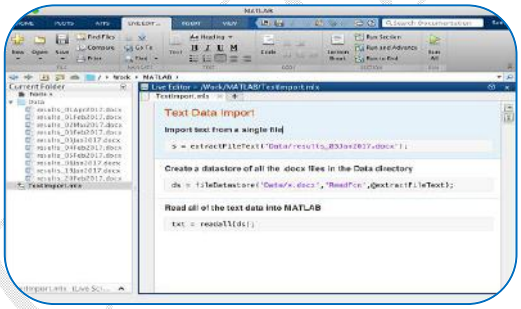
VOLUME - 8 | ISSUE - 6 | MARCH - 2019

INTERACTIVE DATA EXTRACTION ALGORITHM TO EXTRACT DATA FROM MICROSOFT WORD DOCUMENT, APPLICABLE IN CONDUCTING RIVER STUDY

Girish S. Katkar[1] and Dinesh A. Lingote[2]
[1] Department of Computer Science and Application, Art, Commerce & Science College, Koradi, Nagpur, India.
[2] CSIR-National Environmental Engineering Research Institute, Nehru Marg, Nagpur, India.

ABSTRACT :

Microsoft word provides magnificent features in designing document, thus very popularly used for the documentation. Scientific reports involved charts, flowcharts, procedural diagrams, block diagram, Mathematical &Statistical equationsand 3-D diagrams. Microsoft word provides very handy features in generating such documents, hence popularly used in scientific community. Moreover, it provides good features for text designing, tabular data generation, colors and images, thus gives immense pleasure to the user to generation document using this software. In addition to this, designed document can be easily converted into other popular forms like PDF, HTML and rich text format, hence made it worth in using. Microsoft word (.doc) documents are also shared on the web and this is also one media used in disseminating information among the other. If this shared information extracted and re-processed, then this can be very useful information for conducting Research &Development(R&D) activities. This research paper introduces the algorithm developed for extracting data from the word document, which further geo-mapsthe extracted data for the wide dissemination. Considering need to generate data for the KanhanRiver, team has only targeted Kanhan river water quality data for the extraction as team is aimed to generate a central information repository. Using different data generation methodologies central data repository for the Kanhan river is maintained, which can be further utilized for the river and human health management.*



Text Data Import

Import text from a single file

```
s = extractFileText('Data/results_03Jan2017.docx');
```

Create a datastore of all the .docx files in the Data directory

```
ds = fileDatastore('Data/*.docx','ReadFcn',@extractFileText);
```

Read all of the text data into MATLAB

```
txt = readall(ds);
```

KEYWORDS : *Word-Extraction, data generation, extraction, Kanhan River, information system;*

Related work:

Organizations like CPCB (Central Pollution Control Board, Delhi),MPCB (Maharashtra Pollution Control Board, Mumbai) and CSIR-NEERI (Council of Scientific and Industrial Research National Environmental Engineering and Research Institute, Nagpur) periodically conducts water quality study and generates water quality data for the rivers. Dr. S. R. Wate, Former Director of CSIR-NEERI conducted Kanhan river water quality study for three sites namely Kiranapur, Keslapur&Sawangi,and generated water data for the year 2004 and 2005(winter, monsoon and summer). MPCB, Mumbai has setup three water quality monitoring stations on the Kanhan River for monitoring water quality and generate water quality data periodically.

Observing such work and getting inspiration, authors of this research paper has developed an algorithm which can extract data from such reports, the then benefits in conducting various R&D studies on the kanhan river study. In order to generation information for the kanhan river,research team is involved in introducing different possible medium through which information can be extracted.

LBP PUBLICATION

Using such mediums, research team is aiming to generate central repository of the kanhan river water quality data and bio-diversity data for the years' together, so that different research studies can be carried-out for the conservation of river. However, development of central information repository for the Kanhan river using different data generation techniques is unique of its kind.

INTRODUCTION:

The Satpura hills is the widest part of peninsular India, Kanhan River originates from these hills situated in Chhindwara districts of Madhya Pradesh, river progresses to Nagpur district of Maharashtra and Lastly confluences the Wainganga River. To generate central information repository for the kanhan 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. Similarly, to record information Study Points(SPs) are imposed on the river path at 100-meter distance and information is recorded on or closed to SPs (which helps in identifying exact sampling locations). Different data generation methodologies like information extraction, generation and estimation are used to generate data repository.

Research team of CSIR-NEERI had conducted limnobiotic study of the Kanhan River's in 2004 and 2005. The research study was carried out for three sites of the kanhan river and river water quality report was generated [1]. Developed algorithm extracts essential data from such reports, generate central repository and geo-map it for the wide dissemination. Out of mentioned sites, Sawangi is the main point because at this point polluted Nag River confluences the KanhanRiver, hence to monitor KanhanRiver health this point is primarily considered for the data extraction and generation. Figure-1 shows Sawangi SP where Nag River confluences the Kanhan.

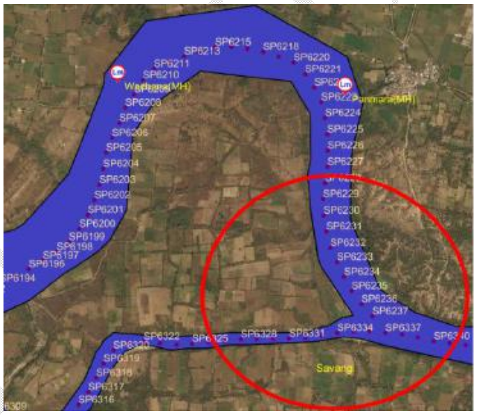
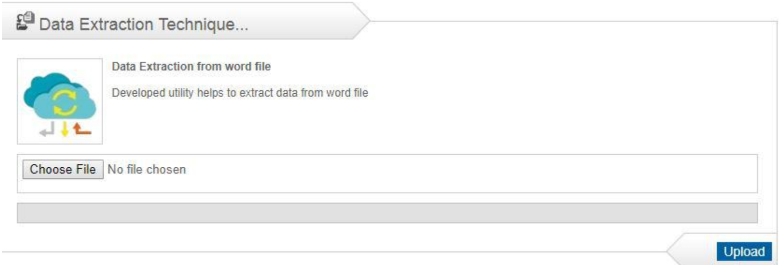


Figure 1. Point where Nag river confluences the Kanhan River

METHODOLOGY:

Java programming language is used to develop this algorithm;Developed Java class files and Java server pages(JSP) are hosted on Tomcat server (8.x). Free available library "org.apache.poi.hssf"[2] is used for reading word document.Google Earth(GE) is used for referral geo-positioning, GE generated files KML (Keyhole Markup language)/KMZ (Keyhole Markup Language Zipped) imported in QGIS(Quantum Geographic Information System). QGIS generated shape files are uploaded in PostGreSQL, PostGresSQL Database wokspace created in Geosever 2.11 and information layers are published through Tomcat web server. The developed algorithm uses following methodology.



Data Extraction from word file

Developed utility helps to extract data from word file

Figure 2. Upload word document into the system folder for extraction

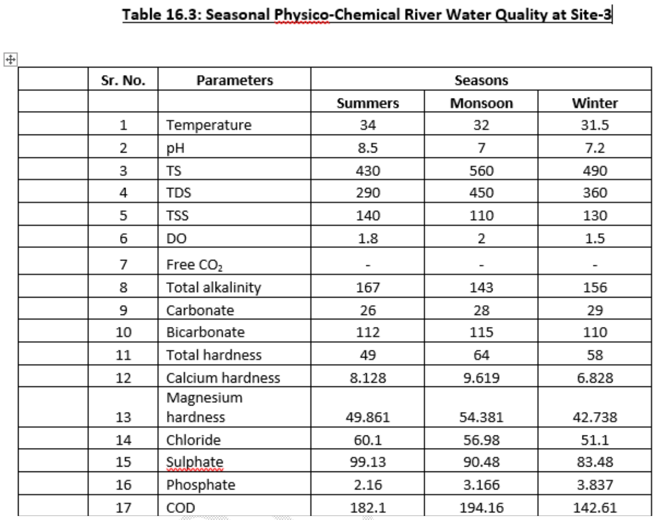
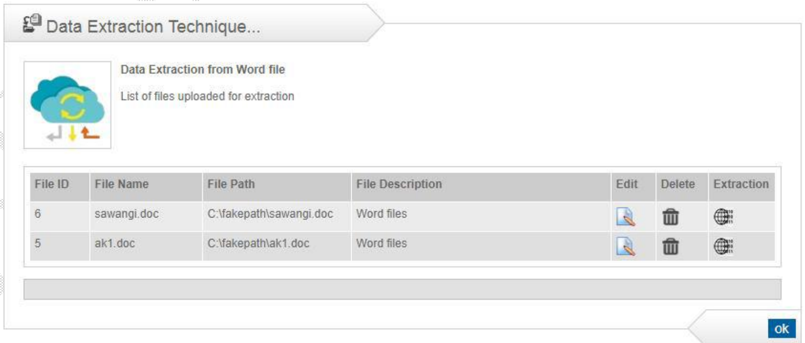


Table 16.3: Seasonal Physico-Chemical River Water Quality at Site-3

	Sr. No.	Parameters	Seasons		
			Summers	Monsoon	Winter
	1	Temperature	34	32	31.5
	2	pH	8.5	7	7.2
	3	TS	430	560	490
	4	TDS	290	450	360
	5	TSS	140	110	130
	6	DO	1.8	2	1.5
	7	Free CO_2	-	-	-
	8	Total alkalinity	167	143	156
	9	Carbonate	26	28	29
	10	Bicarbonate	112	115	110
	11	Total hardness	49	64	58
	12	Calcium hardness	8.128	9.619	6.828
	13	Magnesium hardness	49.861	54.381	42.738
	14	Chloride	60.1	56.98	51.1
	15	Sulphate	99.13	90.48	83.48
	16	Phosphate	2.16	3.166	3.837
	17	COD	182.1	194.16	142.61

Figure 3. Table in Microsoft Word document to be extracted.

As shown in Figure 2, algorithm uploads Microsoft-word(*.doc) file into its system folder.Figure 3 shows Microsoft word document comprised of the table to be extracted [3].Figure4shows the list of word documents uploaded in system folder. Against each uploaded file, extraction icon (given in extraction column) is provided which can be used for applying developed extraction algorithm for the listed files.



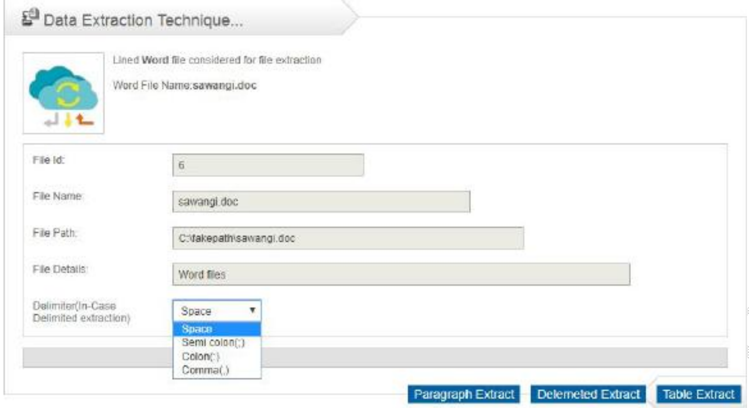
Data Extraction from Word file

List of files uploaded for extraction

File ID	File Name	File Path	File Description	Edit	Delete	Extraction
6	sawangi.doc	C:\fakepath\sawangi.doc	Word files			
5	ak1.doc	C:\fakepath\ak1.doc	Word files			

Figure 4.List of word document uploaded in the system folder for extraction.

Depending on the information comprised in the input word document, algorithm facilitates three extraction techniques like: Paragraph Extract, Delimited Extract and Table Extract (Figure 5).



Data Extraction Technique...

Lined **Word** file considered for file extraction

Word File Name:sawangi.doc

File Id: 6

File Name: sawangi.doc

File Path: C:\fakepath\sawangi.doc

File Details: Word files

Delimiter(In-Case Delimited extraction): Space

Figure 5. Three features for data extraction from the word document.

Paragraph Extract: This is simple method applied for the extraction, Using this method algorithm extract input document line by line till the new-line character (complete paragraph). Subsequently, algorithm reads the document till the end of file and stores output in excel file. This algorithm is applicable when input data is in simple text/paragraph form.

Delimited extract: This extraction method is applicable when input word file is in semi numeric (numbers and text) form. Algorithm uses splitters like: space (" "), comma (,), semicolon (;) and colon(:) to separate input line data from single line to separate columns of excel (Refer Figure 5). Algorithm read input file line by line till the end of file, separates input line by selected splitter, stores into individual column and stores the result in excel file.

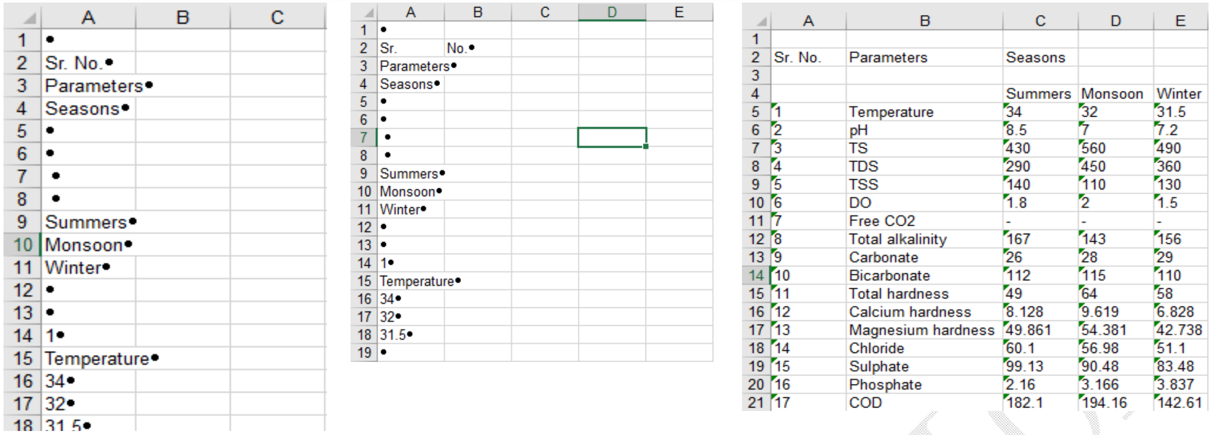
Table Extract: River modeling requires data in numeric and tabular form, hence such data is primarily aimed for the extraction. This extraction option is useful when input document contains numeric and tabular data. Developed extraction algorithm identifies special character, which gets encountered when tabular data comes under the scanner, using this special character algorithm splits the row, separate column and stores in respective columns of the excel. Algorithm reads input stream line by line till the end of file and accordingly generates output in excel file.

Lastly, auto data tuning/correction features of the Excel areutilized for minor data correction, and corrected excel file is imported in PostgreSQL database. PostgreSQL tables are mapped with Geoserver, using such simple techniques extracted data is reflected on the maps and information is disseminated. Various information layers like PH (Potential of Hydrogen), TS (Total Solids), TDS (Total Dissolved Solids), TSS (Total Suspended Solids), DO (Dissolved Oxygen), alkalinity, Carbonate, Bicarbonate, hardness, Magnasium, Chloride, sulphate, phosphate and COD are created and year-wise information is generated on these information layers.

RESULT AND DISCUSSION

Using different data extraction techniques, data is extracted and stored in excel document. As shown in figure 6. we obtained different results, figure 6(a) shows output file obtained when input file extracted using paragraph extraction. Figure 6(b) shows output file obtained using delimited extraction and Figure 6(c) shows output excel file obtained when table extraction algorithm is applied. In this example Table extraction option observed more suitable because input file is in tabular form and table extraction is best suitable option for extracting data from such files. Figure 7 and Figure 8 shows extracted data is imported in postgreSQL and depicted on the map. If observed Figure 1, then significantly no data is recorded, whereas after extraction essential data is generated and accordingly depicted.

LBP PUBLICATION



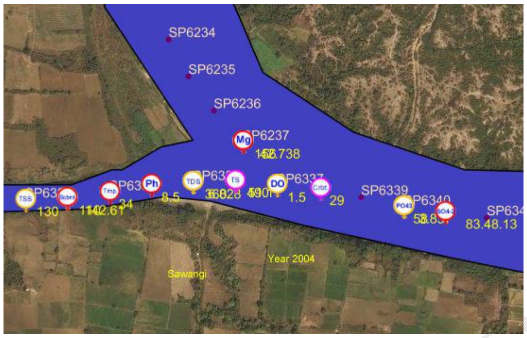
	A	B	C
1	•		
2	Sr. No.•		
3	Parameters•		
4	Seasons•		
5	•		
6	•		
7	•		
8	•		
9	Summers•		
10	Monsoon•		
11	Winter•		
12	•		
13	•		
14	1•		
15	Temperature•		
16	34•		
17	32•		
18	31.5•		

	A	B	C	D	E
1	•				
2	Sr.	No.•			
3	Parameters•				
4	Seasons•				
5	•				
6	•				
7	•				
8	•				
9	Summers•				
10	Monsoon•				
11	Winter•				
12	•				
13	•				
14	1•				
15	Temperature•				
16	34•				
17	32•				
18	31.5•				
19	•				

	A	B	C	D	E
1					
2	Sr. No.	Parameters	Seasons		
3					
4			Summers	Monsoon	Winter
5	1	Temperature	34	32	31.5
6	2	pH	8.5	7	7.2
7	3	TS	430	560	490
8	4	TDS	290	450	360
9	5	TSS	140	110	130
10	6	DO	1.8	2	1.5
11	7	Free CO2	-	-	-
12	8	Total alkalinity	167	143	156
13	9	Carbonate	26	28	29
14	10	Bicarbonate	112	115	110
15	11	Total hardness	49	64	58
16	12	Calcium hardness	8.128	9.619	6.828
17	13	Magnesium hardness	49.861	54.381	42.738
18	14	Chloride	60.1	56.98	51.1
19	15	Sulphate	99.13	90.48	83.48
20	16	Phosphate	2.16	3.166	3.837
21	17	COD	182.1	194.16	142.61

Paragraph extraction (a) Delimited extraction (b) Table extraction (c)

Figure 6. Extracted data using different data extraction technique



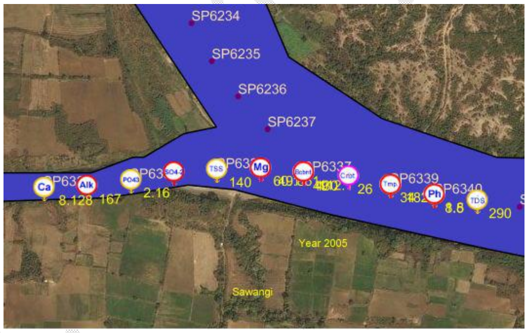


Figure 7.Data extracted and depicted on map, year-2004 **Figure 8.Data extracted and depicted on map, year-2005**

This research paper explains the data extraction from the word document; whereas different data generation methodologies are applied to generate data for large geographical area like rivers. Algorithm also generatescharts to analyze generated data, figure 9 and figure 10 shows data generatedagainst the SPs; shows bar and area graph indicates existence of Potassium inriver for the year 2017. Similarly, Figure 11 and figure 12 shows scattered and spline chart of the sodium concentration in river water for the year 2016.

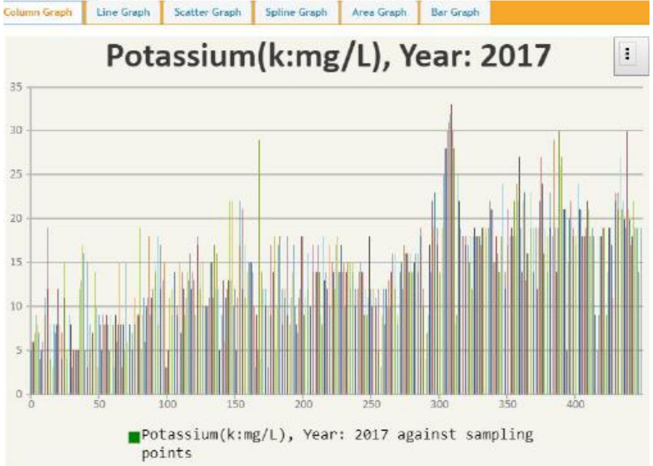


Figure 9. Potassium concentration against SPs for the 2017 – Bar graph

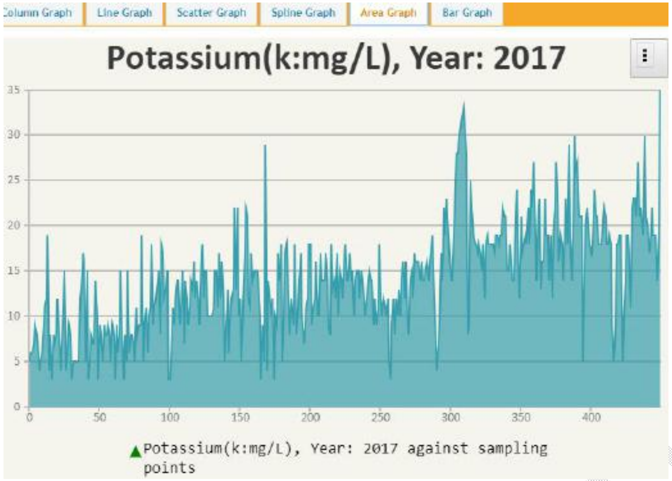


Figure 10. Potassium concentration against SPs for the 2017 – Area graph

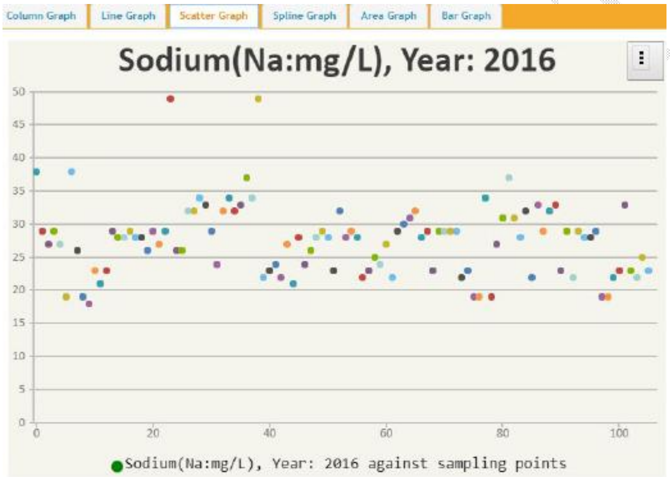


Figure 11. Sodium concentration against SPs for the 2016 – Scattered graph

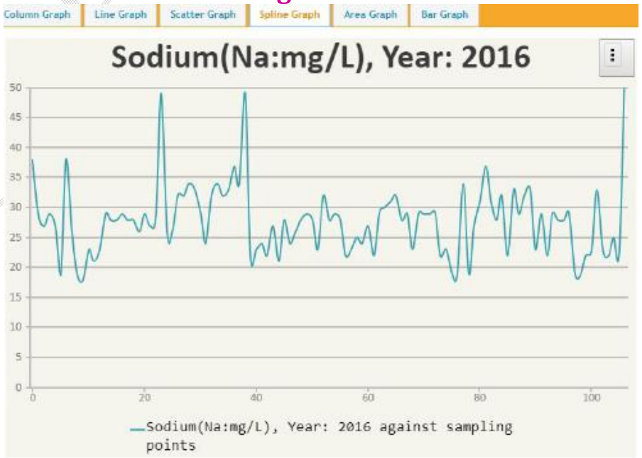


Figure 12. Sodium concentration against SPs for the 2016 – Spline graph

DISCUSSION:

River follows longest path and data generation for such large geographical area is challenging. Considering this, different methodologies are explored for data generation, which can ease the burden of data generation team. Moreover, generated data is geo-mapped and shared with the research team, thus offers great shareware environment to authenticate and validate the generated data. Authenticated data also can be shared on the web, so that common people can easily observe water quality at the required geographical location before its consumption. Using such methodologies, a central repository of the Kanhan River data is generated which can be used for conducting various R&D projects.

CONCLUSION AND FUTURE SCOPE:

River engineering requires tabular and numeric data only. Having such constraint, only numeric structured data is aimed to be extracted. If this algorithm is compared with other data extraction techniques which are applied to extract data from other source like: web/HTML, then library "org.apache.poi.hssf" used for reading Microsoft word document has limitation, as this library is under development and do not provide flexibility in reading and targeting specific data from the input file. Beside, specific tabular data extraction from Web/HTML is quite possible because to extract structured data from Web/HTML specific HTML tags like: <table>, <tr>, <th> and <td> can be targeted. Having such limitation in reading word document, the generated output may contain unwanted data(Figure6(a) and Figure6(b). Therefore, to remove unwanted data from the generated output excel file, available utility of excel for data tuning/editing are used and hence manual intervention is required, thus fully automation in extraction cannot be achieved.

However, if current development is considered then such techniques have wide scope, some notified libraries are under development which can target specific data extraction from the word document, and consequently will be helpful in attaining complete automation in data extracting.

Abbreviations and Acronyms:

R&D	Research & Development
CBCP	Central Pollution Control Board
MPCB	Maharashtra Pollution Control Board
CSIR	Council of Scientific and Industrial Research
NEERI	National Environmental Engineering and Research Institute
HTML	Hypertext Markup Language
GE	Google Earth
KML	Keyhole Markup language
KMZ	Keyhole Markup Language Zipped
SPs	Study Points
JSP	Java Server Pages
QGIS	Quantum Geographic Information System
PH	Potential of Hydrogen
TS	Total Solids
TDS	Total Dissolved Solids
TSS	Total Suspended Solids
DO	Dissolved Oxygen
COD	Chemical Oxygen Demand

REFERENCES

[1.] Data report generated by Dr. S. R. Wate.

[2.] org.apache.poi.hssf is attributed for using free library for reading document.

[3.] 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)

[4.] 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 Sciences and Engineering*, Vol.-6, Issue-12, Dec 2018

[5.] Girish S. Katkar#1, Dinesh A. Lingote*2, Ritesh Vijay@3, "Interactive web-based data generation software applicable for river engineering ", *International Journal of Scientific Research in Computer Science Applications and Management Studies*, Volume 7, Issue 6 (November 2018)

Girish S. Katkar
Department of Computer Science and Application, Art, Commerce & Science College, Koradi, Nagpur, India.

Dinesh A. Lingote
CSIR-National Environmental Engineering Research Institute, Nehru Marg, Nagpur, India.

LBP PUBLICATION