JCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Noise Removal from News Web Sites

N. Narwal

Dept. of Computer Science, Maharaja Surajmal Institute, GGSIP University, New Delhi, India

[*]Corresponding Author: neetunarwal@gmail.com

Available online at: www.ijcseonline.org

Received: 10/Aug/2017, Revised: 21/Aug/2017, Accepted: 18/Sep/2017, Published: 30/Sep/2017

***Abstract*—** Most of the websites comprises of useful information but along with that they contains non-relevant information mostly related to advertisements, copyright, external links etc. This irrelevant information is considered as noise and if we focus on some of the popular English News web sites i.e., Times of India, Hindustan Times, Indian Express etc. consists of 30-40% of news related information and rest are noise content. In this paper we proposed a novel approach that extracts informative content from news web sites in an unsupervised fashion. Our method utilizes the web page segmentation technique to partition the web page into non overlapping rectangular blocks. In our study we used Artificial Neural Network as a classifier to discriminate the rectangular block using their features as relevant or irrelevant blocks. The main content blocks are filtered from the web page and user is presented with clean news web page. Empirical evaluation of our system shows that ANN classifier gives 96.03% accuracy for web content identification that results in accurately filtering of the web page content.

***Keywords*—**Artificial Neural Network, Web Page Segmentation, Visual Blocks, Cosine Similarity

I. Introduction

Websites usually comprises of non-overlapping rectangular blocks related and unrelated to the essence of website, there is a vast amount of nonrelated content considered as noise like advertisement, links to related news, copyright claimant etc. Figure 1 shows a sample of Web page of Hindustan Times uploaded on 8 Aug 2017. Just the visual look of the web page anyone can figure out that the actual news content is almost 40-50% of the complete page and rest of the web page comprises of noise occupies nearly half of the page.

Efficiently extracting informative content from web sites is one of the prominent applications of web content mining. It can be used for the purpose of information retrieval, automatic text categorization, topic tracking, machine translation, abstract summary, helping end users to access the Web easily over constrained devices like PDAs, mobile phones. The mined data content can be used as dataset for numerous real life application like topic related content extraction, spam detection, topic segregation etc. So the web content extraction is widely researched topic has attracted many researchers.

In this paper we present an approach to segregate the real news content from the web site. Then filtered real content are adapted according to the device to present user with a clean web page. In our study we collected our dataset using some of the popular Indian newspaper websites specifically in English.

The identification of the real news content from the webpage is relatively easier for human by using basic intelligence and visual look of website. However it is really difficult for the machine to automatically identify the real content. In this paper we used learning by example and used Artificial Neural Network to classify the rectangular page blocks as real content or noise content.

There are different approaches used by researchers to extract real content page blocks. Based on the techniques used by researchers it is categorized under three main classes:

1) A wrapper is software that extracts the content of web page using algorithm designed for the web page [8]. However the wrapper created for particular web site can’t be used for other websites. The non reusability and complex programming is the limitation of wrappers.

2) Another class of techniques uses web mining techniques, like classification and clustering, to categories or classify the page content of a Web pages [1][2]. The accuracy of these classification and clustering algorithms are better. However most of these techniques are manual or semi-automatic as it needs human intervention and the

complexity of the underlying algorithms is high, so this class of approaches has limited ability for scalable extraction.

3) Third class of techniques uses features in the form of visual and spatial clues of web page to segregate the web page content [3][4]. These approaches can usually perform the extraction in an unsupervised fashion. However most of them rely on some weights or thresholds that are usually determined by some empirical experiments.

We used the third class of technique to identify the web contents based on their visual and spatial features. We used the techniques of learning by example, where Artificial Neural Network classifier is trained on the dataset to identify the main content and noise content of the web page.

The rest of this paper is organized as follows: The Section II outlines related work; Section III, explain the methodology adopted in the system. Section IV, discusses the experiments conducted and results. Section V discusses the application of web content extraction and Finally we presents the conclusion with some final remarks and directions for future work.

II. Related Work

Earlier researchers have suggested some of the database techniques for building a specialized program called wrapper for web information extraction. A wrapper is a specially designed procedure for extracting content from information source and delivers the content of interest in a descriptive representation. A wrapper accepts a query from the user applies a set of extraction rules and returns the relevant information [12]. To extract information from several independent sources, libraries of wrappers are needed. Web is ever changing hence wrappers need to be adjusted dynamically.

The wrappers are generated using manual, semi-automatic and automatic approaches. The manual generation of a wrapper often involves the writing of static code. The developer first understands the structure of the web document and then translates it into program code. The manager of multiple information sources system is the first manual wrapper constructed at Stanford IBM. The goal of the system is to provide methods for accessing data in an integrated fashion from multiple sources, and ensure the information consistency.

Semi-automatic approach makes use of support tools to design the wrapper. Some of these approaches take user input regarding the portion of the web document that need to be extracted. Automatic approach utilizes the power of machine-learning techniques to build wrappers that extract information from the web by analyzing the content without human intervention. A wrapper in this category ranges from simple to relatively complex.

ShopBot [11] is a shopping wrapper, designed to extract information from different web vendors and represent a comparative analysis of a product. The algorithm utilizes the tabular format of the web page content available in the vendor web sites. Information is extracted from different vendor's website by making use of heuristic search, pattern matching and inductive learning techniques.

The Wrapper Induction Environment [19] is a tool for developing wrapper software. It works on structured data in tabular format in the HTML documents. Soft mealy [15] is wrapper software that extracts data from semi-structured web pages using non-deterministic finite automata and an inductive generalization algorithm. It is used to discover the contextual rules from training samples. Gogar et. al. [13] presented self learning wrapper software. The wrapper uses Neural Network to extract information from a previously unseen web page and does not need any site specific initialization. They proposed a method for spatial text encoding that allows encoding of visual and textual content of a web page into a single Neural Network.

Wrapper is a specific approach to extract information from the web, whereas another class of technique namely web page segmentation is a generalized approach of information extraction from the web page. The web page segmentation aims at partitioning a web page into blocks that are cohesive and depict the presentation logic used by the web page designer.

There are different approaches used by researchers for web page segmentation. Some of the most popular approaches are:

1. Fixed length web page segmentation
2. DOM based page segmentation
3. Vision based page segmentation
4. Combined/Hybrid method

The heuristics-based web page segmentation is the most popular technique where researchers have contributed different heuristic strategies and formulated rules for segmentation. Cai D. et. al. [9] used heuristic rules to extract the structural and visual properties of each block of a web page. Their procedure recursively partitions the larger blocks into smaller ones using a top-down approach. However, it is difficult to build generic rules that can be applied on different type of websites. Kohlschutter C. et. al. [17] employed text density as a clue to perform segmentation of a web page.

Chakrabarti D. et al. [10] proposed a graph-theoretic approach for web page segmentation. Considering the segmentation as a minimum cut problem on a weighted graph where the nodes are the DOM tree nodes and the edge weights is the cost of placing the end nodes in the same segment or different segments. A learning based method is used to determine the weights of edges. Different from Cai D. et. al. other researchers Kohlschutter C. et. al. and Chakrabarti D. et. al. obtain a flat segmentation of a web page without knowing the hierarchical structures of the web page.

Kuppusamy K.S. et. al. [18] proposed a model to segment the web page at a fine-grained level and suggested that one should consider only those blocks which contain some informative content. Palekar V.R. et. al. [20], presented an approach that utilizes the visual features of the web pages to perform deep web data extraction.

Kang J. et. al. [16], suggested repetition-based page segmentation (REPS) algorithm which uses the repetitive tag patterns called key patterns from the DOM tree structure of the web page and generate virtual nodes to segment the web page.

Zou J. et. al. [23], proposed html web page segmentation algorithm and applied it on the online medical journal and medical articles. In their methodology they prepared Zone tree by using recursive X-Y cut algorithm and DOM tree analysis along with some other visual features such as background color, font size, and font color etc. The zone tree is then segmented into homogeneous regions.

Cai D. et. al. [9] suggested the vision based page segmentation (VIPS) algorithm for partitioning the web page into blocks. Their study is based on the analysis of the web page and page block relationship using link structure and page layout analysis and they constructed a semantic graph where each node represents a visual block of the web page. Gu X.D. et.al. [14] proposed an automatic top-down tag-tree independent approach to detect the web content structure by simulating the web page layout based on vision. Swezey et. al. [22] proposes a web page segmentation algorithm based on title blocks. The web pages is divided into minimum number of blocks and are classified based on the features of the block as title or non-title block. The title blocks are used as separator and non-title blocks are assembled into web content blocks.

Safi et. al. [21] proposed a framework to enhance the web accessibility for visually impaired people by providing them a first glance web page overview. They suggested a hybrid segmentation algorithm to provide easy navigation of web page. They transformed the layout of the web page into a coarse grain structure, which is transformed into vibrating pages using graphical vibro-tactile language.

III. NOISE REMOVAL FROM NEWS WEBSITES

The methodology used in this study comprises of four phases as shown in Figure 1:

- Web Page Segmentation
- Feature Extraction
- Page Block Identification
- Noise Cleaning and Adapting

The first phase accepts the news web page as input and performs top-down parsing of the web page tree structure by using the functions and methods of Document Object Model (DOM) API (Application Programming Interface). DOM API is a set of function to traverse, modify and delete the tree structure of the web page content.

The top down traversal begins from the root node of the web page then child nodes are recursively traversed until it reaches a level where the node size is below the maximum threshold size of the node (25% of the screen space). However, if the size of splitting node reaches below the minimum threshold size (5% of the screen space) then it is merged with its sibling nodes to form a leaf node. The leaf nodes obtained after segmentation of the web page are the non-overlapping blocks [12].

The *Second phase* receives segregated rectangular block from segmentation algorithm. For each rectangular block feature extraction is performed by analyzing the source code of each block and stores the resultant dataset in an XML file. [21]. For the study the features are segregated into five different categories and they are analyzed for their significance in web page block identification. Table I shows the list of Thirty Three features used in this study.

The five broad categories are:
1. Spatial features – They are related to the spatial or positioning information of the rectangular block inside the web page.
2. Formatting features – They represents the formatting style applied on these rectangular blocks in terms of font-size, color etc.
3. Content features & Hyperlink features - Content features are related to information in terms of text, image, hyperlink and table contained inside each rectangular block and Hyperlink features include information related to internal and externals hyperlinks.
4. Embedded features – Embedded features are related to external script code, hyperlink to external file or object reference present inside the rectangular blocks. Typically external objects or non related content like advertisements

are present inside the rectangular blocks may belong to the same domain or another domain.

Table I: Features of the Rectangular Blocks of web page

Sr. No	Feature Name	Sr. No	Feature Name
1	RelBlock_No_of_Extn_Link	18	Block_left
2	Block_No_of_Extn_Link	19	Block_top
3	Block_No_of_Intl_Link	20	RelBlock_no_of_img_Ext
4	RelBlock_No_of_Inlt_Link	21	Block_no_of_img_Extn
5	Block_no_of_iframe_extn	22	Block_height
6	RelBlock_no_of_iframe_ext	23	Block_no_of_img_intl
7	Block_no_of_iframe	24	Block_fontsize
8	Block_width	25	Block_no_of_table
9	Block_no_of_hyperlink	26	Block_no_of_img
10	RelBlock_width	27	Block_fontwt
11	Block_no_of_script	28	Block_no_of_iframe_intl
12	Block_no_of_script_Intl	29	RelBlock_img_intl
13	RelBlock_no_of_Script_Intl	30	Relno_of_iframe_intl
14	Block_No_of_Script_Extn	31	Rel_block_height
15	Block_Link_Length	32	Block_Textlen
16	RelBlock_textlength	33	RelBlock_no_of_table
17	RelBlock_no_of_Script_Ext		

During the third phase of *Block identification,* the rectangular blocks along with their features are accepted as input to this module. We have used Artificial Neural Network as classifer in this work. In order to train the Classifier these Rectangular blocks are manually labeled with their class either as main content, mix content and noise content block. Each block in the training set is represented as pair (x, y), where x is set of features of the block and y is the class.

Artificial Neural Networks (ANN) is one of the best machine-learning algorithms for solving problems that can’t be solved using conventional algorithm. When the new input is provided to the ANN model, it produces an output similar to the closest matching training input pattern [12]. In neural network model architecture, each node at input layers receives input values, assigns weight to each node and forwarded it to the next layer. The model is trained with the 80% of the dataset.

The key feature of neural networks is that it learns the input/output relationship through training. The response of the neural network is reviewed and the configuration is refined until the analysis of the training data reaches a satisfactory level. In the current system neural network receives 21 inputs and gives 2 outputs with two intermediate layers.

In the last phase of Noise Content filtering, the visual blocks marked as main contents are filtered from the rest of the visual blocks. The main content blocks are then rearranged based on the device dimension and visual block size to fit inside the browser window. The user is presented with the clean web page comprising of pure news content without any noise elements present inside it.

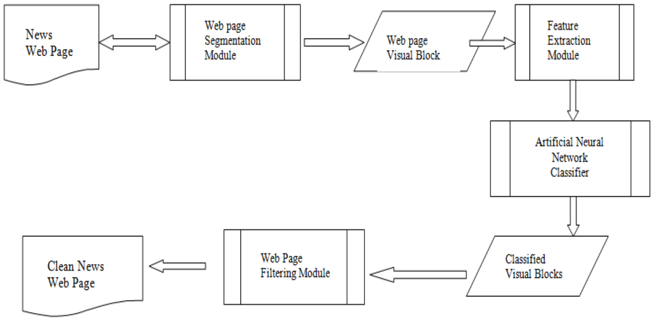

News Web Page
Web page Segmentation Module
Web page Visual Block
Feature Extraction Module
Artificial Neural Network Classifier
Classified Visual Blocks
Web Page Filtering Module
Clean News Web Page

Figure 1: Methodology used in the study

IV. EXPERIMENT

Experiment is conducted on the dataset prepared with 550 news web pages from 50 different news web sites giving total 1500 visual blocks. These blocks are then manually labelled as pure main content, pure noise content and mix of noise and main content.

We implemented the model using feed forward Artificial Neural Network classifier. We evaluated the results using different evaluation measures such as Accuracy, Precision, Recall, F-Measure.

To derive the identification of each block, we have used the approach of learning by example, where the dataset is manually pre labeled with class and trained to build a model. Each block is represented as (x, y), where x is set of similarity measure of each block and y is the class. To test the classifier predictive capability evaluation measure is computed using confusion matrix as shown in Table II.

The confusion matrix is the table of size m by m where m is the total number of class in the dataset, where each row depicts the actual outcome or class given by the classifier and each column depict predicted outcome or class.

True Positive (TP) and True Negative (TN) are indicators of correctness of the classifier. Whereas, True Negative (TN) and False Positive (FP) are indicators of error or mislabelled tuples [11].

Table II Confusion Matrix

		Predicted Class		**Total**
		Yes	No	First
Actual Class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P’	N’	P+N

The accuracy of classifier is the percentage of test tuples that are correctly classified by the classifier.

$$Accuracy = \frac{(TP+TN)}{(P+N)} \quad (1)$$

Precision is a measure of exactness means percentage of tuples labelled as positive.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

Recall is a measure of completeness means percentage of positive tuples labelled as positive.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

F-measure is a combination of precision and recall.

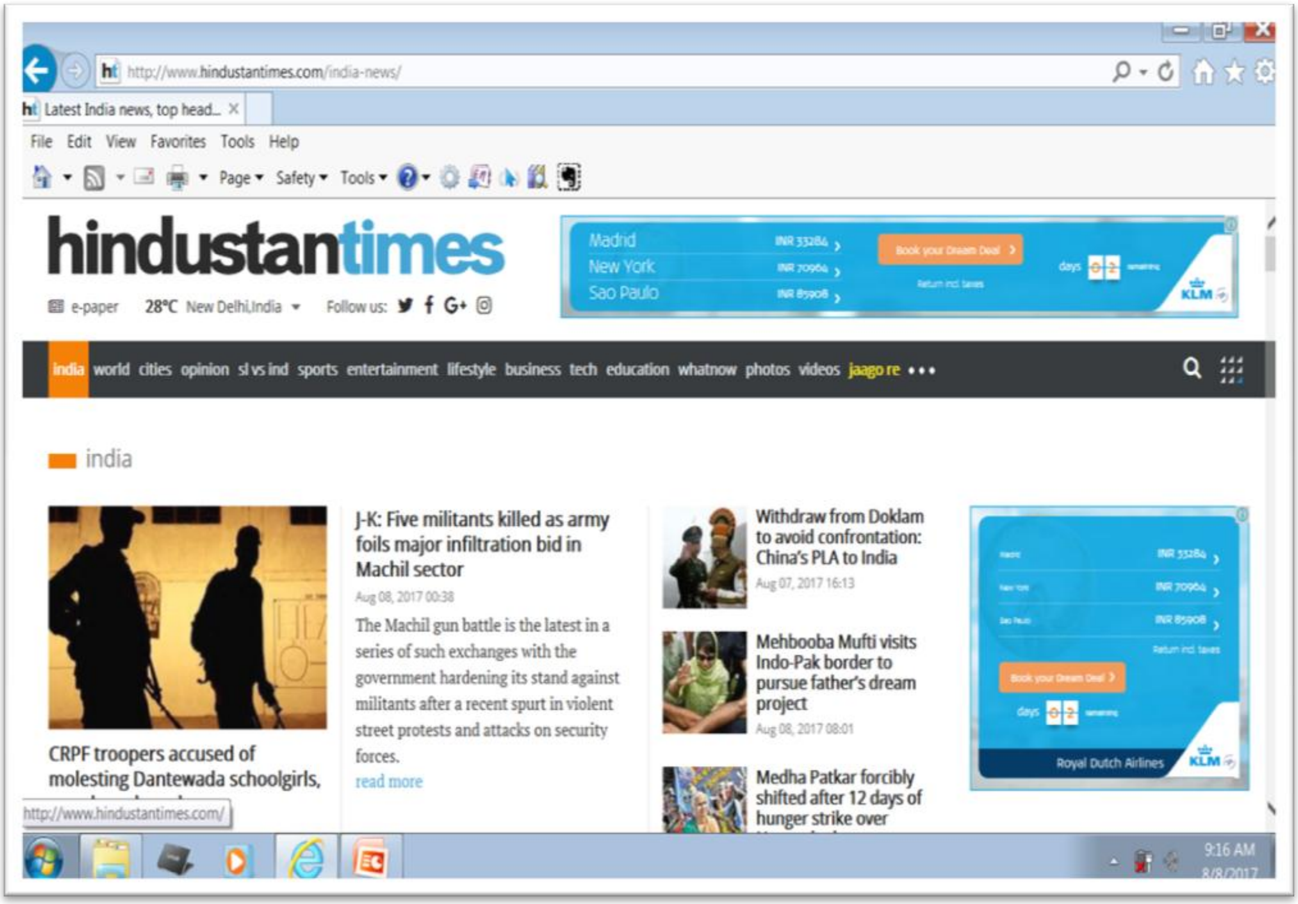
$$F - Measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision+Recall)} \quad (4)$$

We have used feed forward Artificial Neural Network, where the input layer has thirty three neurons and output layer has three neurons. The sigmoid activation function is used to train the model and performance is evaluated after performing five-fold cross validation. Table III shows the efficiency of the classifier depicted in terms of evaluation measures.

Table III. Accuracy Measure of Classifier

Feature set	Accuracy	Precision	Recall	F-Measure
Feed Forward Artificial Neural Network	0.9603	0.8832	0.9295	0.9056

The result depicts that tool provide considerable results in terms of classification of block type and hence can be used for informative content filtering for providing news web page user a clean pure news content.


http://www.hindustantimes.com/india-news/
Latest India news, top head...
File Edit View Favorites Tools Help
Page Safety Tools
hindustantimes
e-paper 28°C New Delhi,India Follow us:
Madrid INR 33284
New York INR 70964
Sao Paulo INR 85908
Book your Dream Deal
Return incl. taxes
days 0 2 remaining
KLM
india world cities opinion sl vs ind sports entertainment lifestyle business tech education whatnow photos videos jaago re •••
india
Royal Dutch Airlines
http://www.hindustantimes.com/
9:16 AM
8/8/2017

CRPF troopers accused of molesting Dantewada schoolgirls,

J-K: Five militants killed as army foils major infiltration bid in Machil sector

Aug 08, 2017 00:38

The Machil gun battle is the latest in a series of such exchanges with the government hardening its stand against militants after a recent spurt in violent street protests and attacks on security forces.

read more

Withdraw from Doklam to avoid confrontation: China's PLA to India

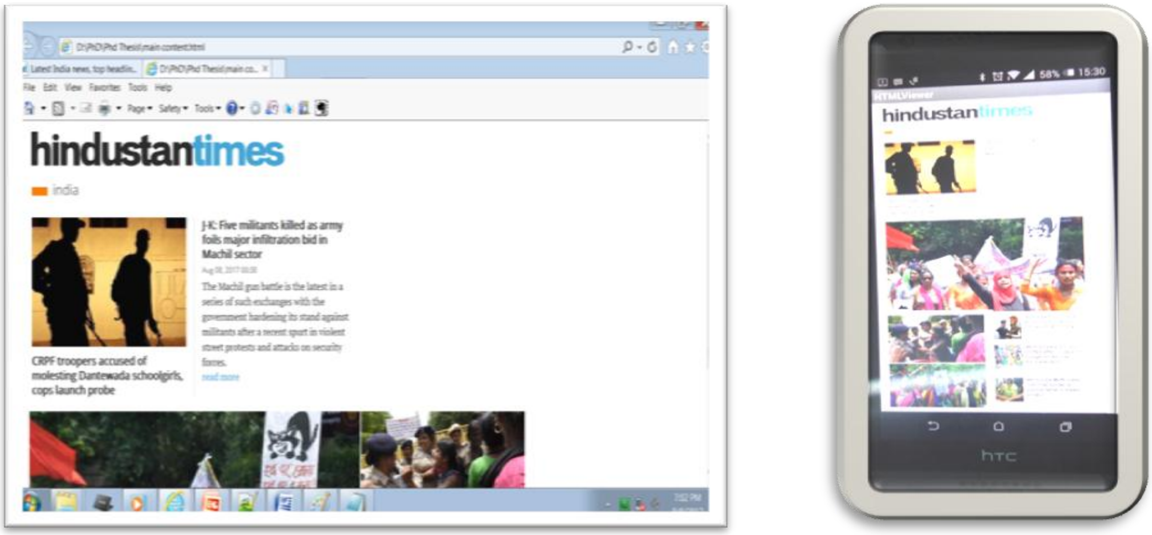
Aug 07, 2017 16:13

Mehbooba Mufti visits Indo-Pak border to pursue father's dream project

Aug 08, 2017 08:01

Medha Patkar forcibly shifted after 12 days of hunger strike over

(a)



D:\PhD\Phd Thesis\main content.html
Latest India news, top headlin...
D:\PhD\Phd Thesis\main co...
File Edit View Favorites Tools Help
Page Safety Tools
hindustantimes
india
HTMLViewer
hindustantimes
htc

J-K: Five militants killed as army foils major infiltration bid in Machil sector

Aug 08, 2017 08:38

The Machil gun battle is the latest in a series of such exchanges with the government hardening its stand against militants after a recent spurt in violent street protests and attacks on security forces.

read more

CRPF troopers accused of molesting Dantewada schoolgirls, cops launch probe

(b) (c)

Fig 2:a) Web page of Hindustan Times web site showing presence of advertisements, external links, internal links with main content uploaded on 8th Aug 2017 (b) Clean News Web Page on Browser Window (c) Clean News Web Page on Mobile Phone

V. Application of Block Filtering System

The block filtering system plays a significant role in various web applications. The output of the model can be utilized for web content personalization, content segregation, search engine crawlers, viewing the web page on small screen device etc.

Block identification can be utilized for topic specific search where user is interested in finding the useful content related to any topic from different web site. The main content from different web sites can be clubbed and displayed to the user.

Another useful application of block identification is displaying selective content of web site on small screen devices. Due to limited screen space, main content and internal links information is sufficient to be displayed to the user.

Conclusion

In this paper we presented the news web page content filtering system that extracts main news content from the web page. We also designed and tested the tool using training data sets. From the experimental results we conclude that the tool provides high precision in classification of informative and non-informative content of the web page and hence is suited for segregating and informative content filtering.

References

[1] C.-N. Ziegler and M. Skubacz, *“Content extraction from news pages using particle swarm optimization on linguistic and structural features,”* in WI ’07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC, USA: IEEE Computer Society, 2007, pp. 242–249.

[2] J. Gibson, B. Wellner, and S. Lubar, *“Adaptive web-page content identification,”* in Proceedings of the 9th annual ACM international workshop on Web information and data management. ACM New York, NY, USA, 2007, pp. 105–112.

[3] J. Prasad and A. Paepcke, “*Coreex: content extraction from online news articles,*” in CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management. New York, NY, USA: ACM, 2008, pp. 1391–1392.

[4] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren, “*Automating content extraction of html documents,”* World Wide Web, vol. 8, no. 2, pp. 179–224, 2005.

[5] I. Muslea, S. Minton, and C. Knoblock, “*A hierarchical approach to wrapper induction,”* in AGENTS ’99: Proceedings of the third annual conference on Autonomous Agents. New York, NY, USA: ACM, 1999, pp. 190–197.

[6] Jaiwei Han, Micheline Kamber, *Data Mining Concepts and Techniques*,Third Edition, ELSEVIER,2012.

[7] Neetu Narwal, Mayank Singh, *Web Content Extraction A Heuristic Approach*, International Journal Of Computer Science and Information Security, Vol 11, No1 , 2013.

[8] N Narwal, S K Sharma, Amit Prakash Singh, *Entropy based content filtering for Mobile Web Page Adaptation*, Proceeding WCI '15 Proceedings of the Third International Symposium on Women in Computing and Informatics Pages 588-594 , ACM New York, NY, USA ©2015 , table of contents ISBN: 978-1-4503-3361-0.

[9] Cai D., Yu S. and Wen J. R., *VIPS: a Vision-based Page Segmentation Algorithm*, Microsoft Technical Report (MSR-TR-2003-79), 2003.

[10] Chakrabarti D., Kumar R., and Punera K., *A graph-theoretic approach to webpage segmentation*, Proceedings of 15th International Conference on World Wide Web, 2008, ACM, pp 377–386.

[11] Doorenbos R.B., Etzioni O., Weld D.S., *A Scalable Comparison-Shopping Agent for the World Wide Web. Technical report UW-CSE-96-01-03*, University of Washington, 1996.

[12] Eikvil, *Information Extraction from World Wide Web - A Survey* , Technical Report 945, Norvegian Computing Center, 1999.

[13] Gogar Tomas, Hubacek Ondrej, Sedivy Jan, *Deep Neural Networks for Web Page Information Extraction, Artificial Intelligence Applications and Innovations,* IFIP Advances in Information and Communication Technology, Vol. 475, 2016, pp 154-163.

[14] Gu X.D., Chen J., Ma W.Y. and Chen G.L., *Visual Based Content Understanding towards Web Adaptation,* Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems , Springer, 2002, pp 164-173.

[15] Hsu C. H. and Dung M. T., *Generating Finite State Transducers for semi structured Data Extraction from the Web*. Information Systems, Vol.23, No. 8, 1998, pp 521-538.

[16] Kang J. , Yang J., Choi J. , *Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices,* IEEE Transactions on Consumer Electronics, Vol. 56, Issue 2, May 2010, pp 980-986.

[17] Kohlschutter C. and Nejdl W. *A densitometric approach to Web page segmentation.* Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp 1173–1182.

[18] Kuppusamy K.S., Aghila G., *A Personalized Web Page Content Filtering Model Based On Segmentation,* International Journal of Information Sciences and Techniques (IJIST) Volume 2, Issue 1, 2012, pp 41-51.

[19] Kushmerick N., Weld D.S., Doorenbos R., Wrapper Induction for Information Extraction. Ph.D. Dissertation, University of Washington. Technical Report UW-CSE-97-11-04, 1997.

[20] Palekar V.R., Ali M. S. And Meghe R., *Deep Web Data Extraction Using Web-Programming-Language-Independent Approach*, Journal of Data Mining and Knowledge Discovery, 2012, pp 69-73.

[21] Safi Waseem, Maurel Facrice, Routoure Jean Marc, Beust Pierre, Dias Gael, *A Hybrid Segmentation of Web Pages for Vibro-Tactile Access on Touch Screen Devices,* . 3rd Workshop on Vision and Language (VL 2014) associated to 25th International Conference on Computational Linguistics (COLING 2014), Aug 2014, pp.95 – 102.

[22] Swezey Robin M.E., Shiramatsu Shun, Ozono Tadachika, Shintani Toramatsu, *Web Page Segmentation Method by using Headlines to Web Contents as Separators and its Evaluation,* International Journal of Computer Science and Network Security (IJCSNS), 2013, Volume 13 Issue 1, pp 1-6.

[23] Zou J., Le D., Thoma G. R., *Combining DOM Tree and Geometric Layout Analysis for Online Medical Journal Article Segmentation*, Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, 2006, pp 119 – 128.

Authors Profile

Dr. Neetu Narwal is Doctorate in Computer Science from Banasthali Vidyapith, Rajsathan, India. She is working as Assistant Professor in Department of Computer Science, Maharaja Sruajmal Institute. She has 16 years of Teaching experience and five years on Industry Experience. Her research areas include web content mining, social media mining.