Web Content Filtering using Machine Learning Approach

Neetu Narwal, Asst. Prof.
Maharaja Surajmal Institute
neetunarwal@gmail.com

***Abstract*—** We see Internet pages as a group of incoherent information presented together as single unit rather than a single cohesive block of information. Most of e-Newspaper websites consist of 30-40% of news related information and rest are the advertisements, link to external websites, copyright information etc. User finds it arduous to focus on news content, and many at times it becomes disturbing. There are many search-oriented applications such as topic specific search application, filtering of actual textual content from surrounding page clutter. In this paper we proposed a novel approach that extracts real content from new web pages in an unsupervised fashion. Our method utilizes the web page segmentation technique to partition the web page into incoherent visual blocks. Artificial neural network is used as classifier to discriminate the visual block based on their features. 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

Web news page can be viewed as a non-coherent group of unrelated news items along with huge amount of noise like advertisement, links to related news, copyright claimant etc. Figure 1 shows a sample of Web news page. In this page, the actual content of news covers almost 30-40% of the complete page and the noise occupies nearly half of the page. Efficiently extracting high-quality content from Web news page is crucial for many Web applications such as 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 extracted content can become the basic data for the further analysis. Various researchers have focused on Content extraction from e-Newspaper websites. [1][2][3][4][5] [6][7].

In this paper we present an approach that extract the main content from the news web pages. We mainly deal with Indian newspaper websites written in English. Since, our results will be used by other real life application hence the accuracy of information identification is important. The identification of the actual content of news in the Web page is a relatively easier for human by making use of his intellect and seeing the visual clues. However it is really difficult for the machine to automatically identify the real content.

There have been many approaches existed to extract content from Web page [8]. These approaches can be divided into three categories based on the techniques used:

1) A wrapper can be generated by wrapper induction system for web content extraction [8]. However the wrapper generated for one web site can't be used for other, hence it becomes website specific. This is a major limitation of wrapper.
2) Some approaches use web mining techniques, such as classification and clustering, to extract content from Web news page such as [1][2]. The accuracy of these systems is better. However most of these techniques rely on human intervention and the complexity of the underlying algorithms is high, so this class of approaches has limited ability for scalable extraction.
3) Some approaches extract content from Web page based on statistics such as [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 approach where the web contents are identified 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 next section outlines related work; In section III, we will provide the details of our system; In section IV, our experiments and results will be discussed; Section V concludes with some final remarks and directions for future work.

Fig 1: Web page of The Times of India web site showing presence of advertisements, external links, internal links with main content uploaded on 4th June 2016.

II. Related Work

Web page can be viewed as collection of non-overlapping blocks and can be easily identified as main content, advertisement, navigation panel, copyright information etc. Earlier researches on web content categorization showed that the block analysis can be used in information retrieval tasks such as searching, classification and clustering. Researchers have worked in the area of Web content extraction using web mining techniques basic three approaches have been studied.

A. Wrapper induction

Chang et. al. [13] presented a detailed survey on the major Web data extraction approaches and compares them in three Dimensions: the task domain, the automation degree, and the techniques used. Some of the well known wrappers developed by researches using supervised method are: WIEN [14], STALKER and SoftMealy [15]; semi-supervised method are : IEPAD [16] and OLERA [17]; and unsupervised method are Dela [18], RoadRunner [19], and EXALG [20].

B. Using Web mining techniques

Many researchers have used web mining techniques for content extraction. Ziegler et. al.[1] presented an approach to extract real content from Web news pages using a particle swarm optimizer (PSO). Gibson et. al. [2], presented another approach for identifying content from a Web page using a sequence labeling technique. The content of a Web page is identified by using a Conditional Random Field sequence labeling model. Reis et. al.[5] in their work used traditional hierarchical clustering techniques to extract the desired news from the Web news sites. Mc Keown et. al. [6] presented an article on content extraction using machine learning program Ripper.

C. Based on Statistics

Content extraction can also be performed based on statistics. Lin et. al. [7] proposed an approach to partition a Web page into several content blocks according to HTML tables, and discovered informative content blocks using statistics based on the occurrence of the features (terms) in the set of pages. Gupta et. al. [4], used DOM tree to navigate the page recursively for content extraction. They applied a series of filtering techniques to remove and adjust specific nodes and leave only the content. The filters are based on statistics on some features of nodes such as link-to-text ratio. Prasad et. al. [3] developed a heuristic technique CoreEx for extracting the main article from Web news pages. CoreEx traverse the DOM tree of the page and scores every node based on the amount of text, the number of links it contains and additional heuristics.

III. News Web Page Filtering System

The first phase, we takes web page as input and parse the tree structure using top-down approach by using the functions and methods of Document Object Model (DOM) API (Application Programming Interface). DOM API used to traverse, modify and delete the tree structure of the web page elements even at run time. In the top down approach the 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].

In the *Second phase* of the features are obtained after analyzing the web page blocks [21]. These features are segregated into five different categories and analyzed their significance for web page block identification. The categories are:

1. Spatial features – it is related to the positioning information of the visual block with reference to the web page.

2. Formatting features – it represents the formatting style applied on the visual block.

3. Content features & Hyperlink features - Content features are related to information in terms of text, image, hyperlink and table contained inside the block.

4. Embedded features - Web page usually includes few external objects that are embedded in the web content. These embedded objects may belong to the same domain or another domain.

In the *Block identification phase* the features extracted from second phase is used for classification. We used learning by example approach where the dataset is manually pre labeled with class and trained to build a model. Each block is represented in pair (x, y), where x is set of features of the block and y is the class.

In this work we used feed forward Neural Network (ANN) to train our model. 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, perform processing and send it to the next layer.

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, the content marked as main contents are filtered from the list of visual blocks. These blocks are rearranged accordingly to fit the browser window. The news web page user is presented a clean news web page.

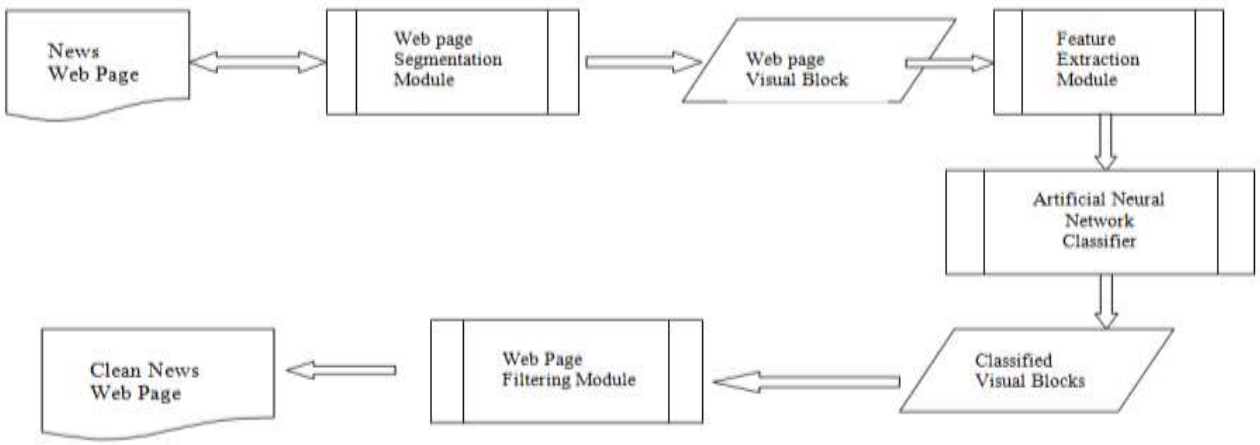

Fig 2. Methodology used in the study

IV. EXPERIMENT

Experiment is conducted on the dataset prepared with 350 news web pages from 40 different web sites comprising of five different categories i.e., science, academics, fiction, sports and news, giving total 800 visual blocks. These blocks are then manually labeled 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. In this paper we have used Artificial Neural Network (ANN) techniques to train the model.

To test the classifier predictive capability evaluation measure is computed using confusion matrix as shown in Table I. 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 mislabeled tuples [11].

Table 1. 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 labeled as positive.

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

Recall is a measure of completeness means percentage of positive tuples labeled 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 neural network, where the input layer has three neurons and output layer has two neurons. The sigmoid activation function is used to train the model and performance is evaluated after performing five-fold cross validation. Table II shows the efficiency of the classifier depicted in terms of evaluation measures.

Table 2.Accuracy Measure of Classifier

Feature set	Accuracy	Precision	Recall	F-Measure
Feed Forward 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.

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 developed a tool for this. We tested the tool through experiment using the 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] D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender, "Automatic web news extraction using tree edit distance," in WWW '04: Proceedings of the 13th international conference on World Wide Web. New York, NY, USA: ACM, 2004, pp. 502–511.

[6] K. Mc Keown, R. Barzilay, J. Chen, D. Elson, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman, "Columbia's newsblaster: new features and future directions," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4. Association for Computational Linguistics Morristown, NJ, USA, 2003, pp. 15–16.

[7] S.-H. Lin and J.-M. Ho, "Discovering informative content blocks from web documents," in KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2002, pp. 588–593.

[8] 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.

[9] Hakan Cevikalp, Member, IEEE, Diane Larlus, Matthijs Douze, and Frederic Jurie, Member, IEEE, Local subspace Classifiers : Linear and Non Linear Approaches, IEEE Transactions, 2007.

[10] Stephen Grossberg Non Linear Neural Networks : Principles, Mechanisms and Architectures, Neural Networks, Pergammon Journal, Vol 1 pp 17-61, 1988.

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

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

[13] C.-H. Chang, M. Kayed, R. Girgis, and K. Shaalan, "A survey of web information extraction systems," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 10, pp. 1411–1428, Oct. 2006.

[14] N. Kushmerick, "Wrapper induction for information extraction," Ph.D. dissertation, 1997, chairperson-Daniel S. Weld.

[15] C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semi-structured data extraction from the web," *Inf. Syst.*, vol. 23, no. 9, pp. 521–538, 1998.

[16] C.-H. Chang and S.-C. Lui, "Iepad: information extraction based on pattern discovery," in *WWW '01: Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM, 2001, pp. 681–688.

[17] C.-H. Chang and S.-C. Kuo, "Olera: semisupervised web-data extraction with visual support," *Intelligent Systems, IEEE*, vol. 19, no. 6, pp. 56–64, Nov.-Dec. 2004.

[18] J. Wang and F. H. Lochovsky, "Data extraction and label assignment for web databases," in *WWW '03: Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA: ACM, 2003, pp. 187–196.

[19] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards automatic data extraction from large web sites," in *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 109–118.

[20] A. Arasu, H. Garcia-Molina and S. University, "Extracting structured data from web pages", in SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2003, pp. 337–348.

[21] 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 doi> 10.1145/2791405.2791470 .