Web Content Extraction by Integrating Textual and Visual Importance of Web Pages

K. Nethra
Pg Scholar (CSE),
Sri Ramakrishna Engineering College,
Coimbatore.

J. Anitha
Assistant Professor (IT) Sl. Gr,
Sri Ramakrishna Engineering College,
Coimbatore.

ABSTRACT
A Web page has huge information and the information in the Web pages is useful in real world applications. The additional contents in the Web page like links, footers, headers and advertisements may cause the content extraction to be complicated. Irrelevant content in the Web page is treated as noisy content. A method is necessary to extract the informative content and discard the noisy content from Web pages. An integration of textual and visual importance is used to extract the informative content from Web pages. Initially a Web page is converted in to DOM (Document Object Model) tree. For each node in the DOM tree, textual and visual importance is calculated. Textual importance and visual importance is combined to form hybrid density. Density sum is calculated and used in content extraction algorithm to extract the informative content from Web pages. Performance of Web content extraction is obtained by calculating precision, recall, f-measure and accuracy.

Keywords
Web Content Extraction, Web content Mining, DOM tree, Vision based Page Segmentation.

1. INTRODUCTION
Huge growth of internet has made World Wide Web as a significant place for distributing and gathering information [13]. Mining the data on the Web has become a major task for locating useful information from the Web. The information's that are considered as useful information on the Web is usually polluted by huge amounts of noise data's such as navigation bars, advertisements, notices etc.

Performance of Web mining can be improved by identifying and discarding noises from Web pages. Effective Web data extraction requires main content (i.e. without irrelevant data) to be gathered, processed and warehoused quickly. The irrelevant data is to be removed properly and hence the main content is considered as the informative part (i.e. Topics related on the Web page) which can provide some useful information to the reader.

Content extraction techniques that extract the main content discarding unwanted data is useful for many real-world applications [7]. Numerous content extraction techniques gave poor performance since they were not able to adapt to these changes because particular HTML tags (e.g., fonts, <table> and <td>) that are utilized before are not included in recent Web pages.

In this paper, a content extraction algorithm that combines textual and visual importance to gather main content from the Web pages is proposed [4]. DOM tree construction is done using a DOM parser giving the Web page content as input which classifies the nodes (i.e. tags) using a Parent-Child relationship.

For each node, Textual information is observed through the measures - Text Density and Composite Text Density. Maximum value for text density indicates that the node represents the main content within a Web page whereas a minimum value indicates the noise. Composite Text Density calculation is done by including information about the hyperlinks. Also Visual information [12] is calculated through the visual measure - VIPS algorithm [6] which splits the Web page into blocks and calculates with the display position and size of each block in a Webpage. Finally the above stated measures are integrated and content extraction algorithm is used to gather the main content from the Web pages. The advantage is that no assumptions are made about the organization of an input Web page or about the arrangement of the tags and also the original structure of the Web page are preserved since the operation is performed using DOM tree.

The paper is organized as follows: after concise reviewing related work in Sect. 2, proposed method discussed in Sect. 3, including definitions of text density and composite text density, definition of visual importance and hybrid text density, as well as how to select the threshold and the algorithm to gather the content. Finally, conclusions and plans for future research are discussed.

2. RELATED WORK
Web Content Extraction techniques available are grouped into three major categories – Automatic Extraction, Hand-craft rules, Template detection

2.1Automatic Extraction
Christian Kohlschutter, Peter [10] - Automatic Extraction is the method of extracting the Web page data automatically. Web page segmentation is done based on three approaches - visual-based segmentation, DOM-based segmentation, and location-based segmentation.

Fankhauser, Wolfgang Nejdl [11] proposed an approach for boilerplate detection using Text features where the Text Content of a Web page is grouped into two classes - long text and short text. In systematical analysis words in the short text are removed.

Marco Baroni, Francis Chantree, Adam kilgariff, Serge Sharoff [2] proposed CleanEval as a shared task for cleaning arbitrary Web pages. First the data preparation is done by data selection and annotation is performed with instructions like removal of HTML/Java code or using "boilerplate" method or by basically encoding the structure of the page using minimal set of symbols to mark the different sections in a Web page such as beginning of headers, paragraphs and list elements. At last the Scoring is measured to find the similarity between the two cleaned versions of the same file.

John Gibson, Ben Wellner, Susan Lubar [5] defined a method to identify the target content in a Web page. Content

identification is done by sequence labeling method and boundary detection method. Some of the models used for sequence labeling are Maximum Entropy Markov models (MEMM), Conditional Random Fields (CRF) and Maximum Entropy Classifiers (MaxEnt).

Pinto et al., extended BTE in Document Slope Curves (DSC) method [16]. Using a windowing technique they located many document regions in which word tokens are high often than tag tokens while decreasing the complexity to linear runtime.

Mantratzis et al., presented Link Quota Filters (LQF) [15] that identifies the lists of link and navigation elements. The main plan is to identify a DOM element that has a text in hyperlink anchors.

Gottron gave the idea of Content Code Blurring (CCB) [19] which is established by identifying the regions in the source code character sequences that indicate the text that are formatted similarly.

2.2Hand-craft rules

Hand crafted rule generates rule using string manipulation function. Hand-crafted rules cannot be applied for large number of source sets.

2.3Template detection

Several procedures have been proposed for template detection [14]. First method uses the concept of page let to divide a Web page. A page let takes into account the count of hyperlinks in a HTML element. Template is defined as a page let where its frequency exceeds a threshold value. Second method divides the Web pages based on the Web page language tags like HTML <TABLE> tag. Third Method uses the similar partition procedure as second method. Entropy of a block is computed by extracting the keywords of each block. Templates are defined as blocks with small entropy value.

The stated proposal has two issues:

(1) Accuracy of the method depends on feature selection.

(2) To gather the features of each and every block it utilizes more time.

3. PROPOSED METHOD

Initially a Web page is transformed into DOM tree using DOM parser. The content extraction technique proposed uses a two way approach. The first approach known as Textual Information considers that noise part has short sentences comprised of less number of textual information and main part contains more number of textual information and it also has less number of hyperlinks as compared with the noise. The second approach known as Visual Information [12] considers that the main content is generally displayed in the center of a Web page.

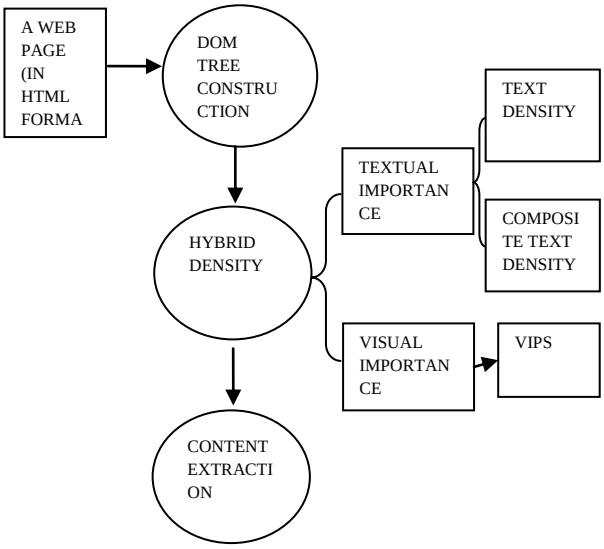

Fig 3.1 Architecture of a hybrid approach

For each nodes in the DOM tree textual and visual importance are calculated. Textual importance of a Web page is acquired by calculating text density and composite text density. Visual importance of a Web page is attained by segmenting a page using VIPS (Vision based Page Segmentation) algorithm and for each block, normal distribution is used to calculate Probability density function. Finally both are integrated to form Hybrid Density. Density Sum technique is also used which extracts integral content from a Web page. Content Extraction algorithm is used to extract the focal content of a Web page.

3.1DOM tree

Document Object Model (DOM) [4] is a platform independent and language independent interface to access or to reform the content, style of the documents and structure. A HTML page is similar to a DOM tree whose detailed text and images are leaf nodes and tags are internal nodes.

3.2Textual Importance

3.2.1Text Density

In order to gather information from a Web page the textual features of both content and noise are taken into consideration [4]. It was found that the noise part has short sentences comprised of less number of textual information and main part contains more number of textual information and it also has less number of hyperlinks as compared with the noise.

A Web page is parsed using a DOM parser which is in turn represented as a DOM tree. The count of characters and tags in each node can be identified. Statistical information can be integrated to the node as follows.

– Char Number (Ci): Count of characters in each sub tree.

– Tag Number (Ti): Count of tags in each sub tree.

A ratio of the count of characters to the count of tags for a node is computed and the Text Density (TDi) is defined as follows:

$$TDi = \frac{Ci}{Ti} \quad (1)$$

Where i is a tag in a Web page, Ci is the number of characters in the range I, Ti is the number of tags in the range i.

If the value of Ti is 0 then by default Ti value is set to be 1. TDi is a estimation of the density of all node's text in a Web page. Initially ahead computation the style tags, the script and comment are discarded from the DOM tree since such information may affect the result.

3.2.2Composite text density

In order to calculate composite text density the number of hyperlinks is considered since noise in a Web page generally consists of hyper links [4]. Additional statistical information per node is calculated as follows:

– Link Char Number (LCi): Count of all hyperlink characters in each sub tree.

– Link Tag Number (LTi): Count of all hyperlink tags in each sub tree.

Composite Text Density calculation is done as follows:

$$CTDi = \frac{Ci}{Ti}\log_{\ln(\frac{Ci}{Ti}LCi+\frac{LCb}{Cb}Ci+e)}\left(\frac{Ci}{LCi}\times\frac{Ti}{LTi}\right) \quad (2)$$

Where i is a tag in a Web page, *Ci* is the count of all characters in the range *i* , *Ti* is the count of all tags in the range *i* , *LCi* is the count of all hyperlink characters in the range *i* , ¬ *LCi* is the count of all non-hyperlink characters in the range i.

3.3Visual importance

Visual importance considers the relative display positions and sizes of blocks for content extraction and is attained by implementing VIPS algorithm [6] and probability density function as follows:

3.3.1VIPS Algorithm

In Vision based Page Segmentation algorithm [6], the vision-based content organization of a Web page is attained by merging the DOM structure and the visual cues. This top-down algorithm performs in three steps: content structure construction, block extraction and separator detection. Initially the Web page is separated into many sizeable blocks and the hierarchical arrangement of this level is stored in a pool. For all big block, the identical segmentation process is iteratively done until small blocks are reached which cannot be segmented further (i.e. Degree of Coherence - DoC values are greater than pre-defined PDoC). Degree of coherence values is assigned according to its intra visual differences.

3.3.2 Probability Density Function

For each block obtained from VIPS algorithm, the relative locations and size are gathered. (Block is denoted as NodeL) and its sided boundaries - xL1 and xL2 (xL2 > xL1) on the horizontal dimension are found[4].

– The area under NodeL is considered significant. The visual importance distribution is illustrated by a standard normal distribution N (0, 1). Points of xL1 and xL2 are mapped to point's −1 and +1 on the variable axis of the standard normal distribution function, respectively.

– Based on the above stated mapping, a normal distribution N (μ, σ2) for the visual importance is generated. The mean of the distribution is assigned as μ = (xL1+xL2)/2 and variance as σ = $(xL2-xL1)^2/2$.

The probability density function is as follows

Normal distribution function equation

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma}} \quad (3)$$

Where f(x) is the probability density function as xL1 and xL2 are horizontal boundaries of the node with a largest horizontal span under the root node.

If i is a leaf node in the DOM tree and the horizontal coordinates of its displaying boundaries are x1 and x2, then its Visual Importance (VIi) is as follows:

$$VIi = \int_{x1}^{x2} f(x) \quad (4)$$

3.4 Hybrid density

Hybrid Density calculation [4] is performed based on the obtained values from textual and visual importance.

If i is a leaf node in the DOM tree, then its Char Number (Ci) is promoted to Hybrid Char Number (HCi) by setting visual importance value as weight.

$$HCi = V\,Ii * Ci \quad (5)$$

For other tag nodes, the Hybrid Char Number is defined as the sum of all its sons Hybrid Char Numbers.

If i is a node in the DOM tree, then its Hybrid Text Density (HT Di) is:

$$HTDi = \frac{HCi}{Ti}\log_{\ln(\frac{HCi}{LCi}LCi+\frac{LCb}{HCb}HCi+e)}\left(\frac{HCi}{LCi}\times\frac{Ti}{LTi}\right) \quad (6)$$

In which all the appearances of Ci are substituted by HCi. Ci and Cb for node i and the <body> tag are updated to HCi and HCb. If Ti is 0, it is set to be HCi /Ci, where Ci and HCi are the initial and the Hybrid CharNumbers, respectively.

3.5 Content extraction

A threshold t is determined that divides nodes into content or noise sections. The best value of the threshold is found and the content is extracted.

Density Sum

In some articles the main content may have abnormally low text density. If it is treated as defined, some content may be lost or some noisy nodes may be retained. Data Smoothing considers the values in the outlying ranges, increases cohesiveness within and between sections. Density Sum is based upon the fact that an information block belongs to an ancestor node in the DOM structure and also the text density of information nodes is much greater than that of noise nodes. The content block will get a maximum value if its children's text densities are added.

If N is a tag under a Web page and i is a child of N, then N's Density Sum is the total of all its children's text densities:

$$DensitySum_N = \sum_{i\in C} Text\ Density_i \quad (7)$$

Where *C* is the set of *N*'s children and *Text Density i* is the text density of tag *i*.

Threshold

The Web page may contain many content block. In these cases the maximum Density Sum tag in the whole page is found without threshold value [4].

Then the minimum text density of the node is set in the path from the maximum Density Sum tag to <body> tag as threshold. Then for each node in the DOM tree whose text density is higher than the threshold, the same method is applied to gather the information.

Content extraction using Density Sum [4]

Algorithm 2 Pseudo code of Extract Content (N)

1 **INPUT**: N

2 **if** N.Text Density >= threshold **then**

3 T ← Find Max Density Sum Tag(N)

4 Mark Content (T)

5 **for** child node C in N **do**

6 Extract Content (C)

7 end for

8 end if

5. PERFORMANCE EVALUATION

Performance is obtained by calculating the metrics. Metrics like precision, recall, F-measure and accuracy are used. The accuracy metric is used to measure the percentage of correct predictions for the overall data.Precision finds the fraction of records which actually turns out to be positive in the group where the classifier has declared as a positive class. Recall finds the fraction of correct instances among all instances that actually belong to a relevant subset. A measure which combines precision and recall is F-measure, which can be also known as the harmonic mean of precision and recall.

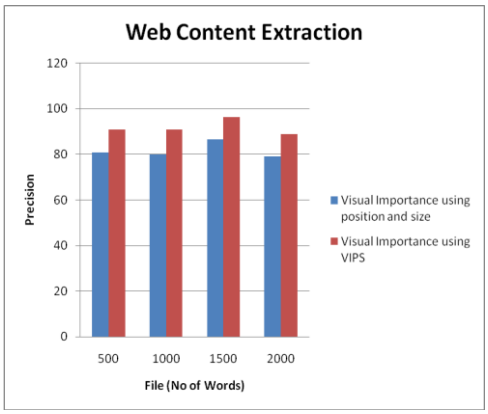

Fig 4.1 Performance comparison based on precision

Fig 4.1, gives the comparison of visual importance using VIPS and visual importance using position and size based on the metric precision. When numbers of words in a HTML file increases, precision of visual importance using VIPS is high when compared with visual importance based on position and size. Based on precision, visual importance using VIPS achieves 89% whereas visual importance using position and size achieves only 81%.

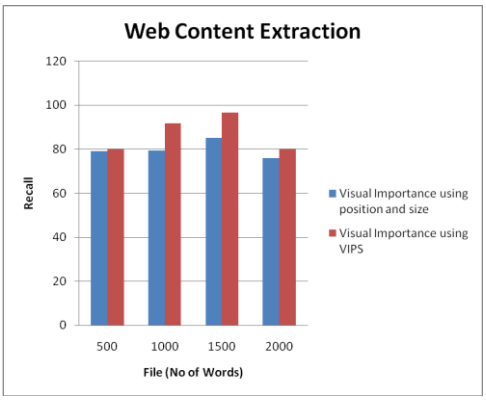

Fig 4.2 Performance comparison based on recall

Fig 4.2, gives the comparison of visual importance using VIPS and visual importance using position and size based on the metric recall. When numbers of words in a HTML file increases, recall of visual importance using VIPS is high when compared with visual importance using position and size. Based on recall, visual importance using VIPS achieves 81% whereas visual importance using position and size achieves only 76%.

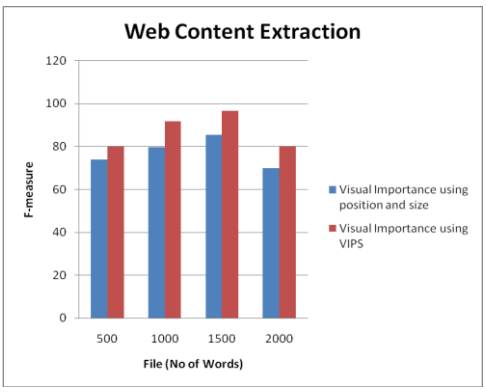

Fig 4.3 Performance comparison based on F-measure

Fig 4.3, gives the comparison of visual importance using VIPS and visual importance using position and size based on the metric f-measure. When numbers of words in a HTML file increases, f-measure of visual importance using VIPS is high when compared with visual importance using position and size. Based on f-measure, visual importance using VIPS achieves 80% whereas visual importance using position and recall achieves only 70%.

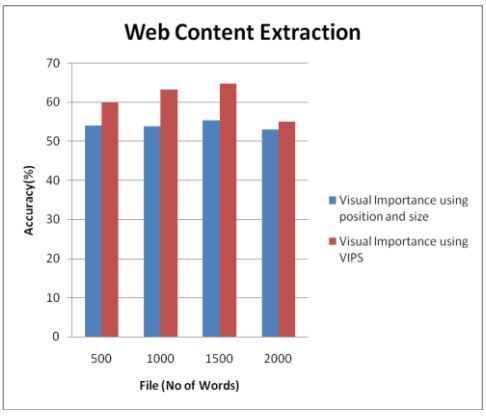

Fig 4.4 Performance comparison based on Accuracy

Fig 4.4, gives the comparison of visual importance using VIPS and visual importance using position and size based on the metrics accuracy. When numbers of words in a HTML file increases, accuracy of visual importance using VIPS is high when compared with visual importance using size and position. Based on accuracy, visual importance using VIPS achieves 55% whereas visual importance using size and position achieves only 53%.

6. CONCLUSION

In this paper a method for extraction of Web content proposed is based on textual and visual importance of a Web page. A Web page is translated to DOM tree and for each DOM nodes, textual importance and visual importance (more efficient VIPS algorithm is used for page segmentation and for each block probability density function) is calculated. Based on the textual and visual importance, a hybrid density is attained and used in content extraction algorithm to gather the main content. Performance is obtained by calculating the metrics like precision, recall, f-measure and Accuracy. Visual importance calculated using VIPS performs better than visual importance calculated using position and size of the each node.

Extraction method proposed can be used for efficient gathering of news from online news articles, Data from digital libraries. In future it can be implemented in web crawler for efficient content extraction. Instead of VIPS, ViDE algorithm can be used under visual importance to improve the accuracy of content extraction.

7. REFERENCES

[1] Baluja, S.(2006).Browsing on smalls screens: Recasting web-page segmentation in to an efficient machine learning framework. In WWW '06: proceedings of the 15th international conference on World Wide Web. NewYork: NY,USA, ACM. pp.33–42

[2] Baroni,M ., Chantree,F. ,Kilgarri,A., Sharo, (2008). Cleaneval : A competition for cleaning web pages. In Proceedings of the sixth international,language resources and evaluation (LREC'08).

[3] Chen,Y., Ma,W.-Y. ,& Zhang,H.-J. (2003). Detecting web page structure for adaptive viewing on small form factor devices. In Proceedings of the12th international conference on World Wide Web (WWW'03). NewYork, NY,USA:ACM. pp. 225–233

[4] Dandan Song, Fei Sun, Lejian Liao." A hybrid approach for content extraction with text density and visual importance of DOM nodes". In the proceedings of Springer Knowl Inf Syst, DOI 10.1007/s10115-013-0687-x, Verlag London 2013.

[5] Debnath, S. ,Mitra,P.,Pal,N.,&Giles,C.L. (2005). Automatici dentification of informative sections of web pages. IEEE Transaction on Knowledge and Data Engineering, 17(9), 1233–1246.

[6] Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma." VIPS: a Vision-based Page Segmentation Algorithm". Technical Report MSR-TR-2003-79, Microsoft Research, 2003.

[7] Finn A, Kushmerick N, Smyth B (2001) Fact or fiction: content classification for digital libraries. In: Joint DELOS-NSF workshop: personalization and recommender systems in digital libraries

[8] Gibson, J.,Wellner,B.,&Lubar,S.(2007). Adaptive web-page content identification. In WIDM '07:Proceedings of the 9th annual ACM international workshop on Web information and data management, New York, NY,USA,ACM. pp. 105–112

[9] Gottron T (2008) Content code blurring: a new approach to content extraction. In: Proceedings of DEXA '08, pp 29–33

[10] Kohlschutter, C(2009).A densitometric analysis of web template content. In WWW 09: Proceedings of the 18th international conference on World Wide Web. New York,NY,USA:ACM.

[11] Kohlschutter,C.,Fankhauser,P,&Nejdl,W (2010). Boiler plate detection using shallow text features. In Proceedings of the third ACM international conference on Web search and datamining (WSDM'10). NewYork ,NY,USA:ACM.pp. 441–450

[12] Kovacevic, M.,Diligenti, M., Gori,M., & Milutinovic,V.(2002).Recognition of common areas in a web page using visual information:A possible application in a page classification.In the proceedings of 2002 IEEE international conference on data mining(ICDM'02),MaebashiCity,Japan,December.

[13] Lan Yi ,Bing Liu,Xiaoli Li."Eliminating Noisy Information in web pages for Data Mining" . In the Proceedings of ACM 1-58113-737-0/03/0008,*SIGKDD .03*, August 24-27, 2003, Washington, DC, USA[1]

[14] Liang Chen, Shaozhi Ye, Xing Li." Template Detection for Large Scale Search Engines".In the proceedings of ACM 1-59593-108-2/06/0004SAC'06 April 23-27, 2006, Dijon, France.[3]

[15] Mantratzis C, Orgun M, Cassidy S (2005) Separating XHTML content from navigation clutter using DOM-structure block analysis. In: Proceedings of HYPERTEXT '05, pp 145–147

[16] Pinto D, Branstein M, Coleman R, CroftWB, King M, LiW,Wei X(2002) QuASM: a system for question answering using semi-structured data. In: Proceedings of JCDL '02, pp 46–55

[17] Uzun Erdinc,Hayri Volkan Agun ,Tarik Yerlikaya.(2013).A hybrid approach for extracting informative content from web pages. In the Proceeding of Elsevier journal.

[18] Uzun E.,Yerlikaya,T. , & Kurt, M. (2011b). A light weight parser for extracting useful contents from web pages. In 2nd International symposium on computing in science&engineering–ISCSE2011,Kusadasi, Aydin,Turkey,pp.66–72.

[19] Weninger T, Hsu WH, Han J (2010) CETR—content extraction via tag ratios. In: Proceedings of WWW'10. NY, USA, New York, pp 971–980.

[20] Yves Weissig, Thomas Gottron."Combinations of Content Extraction Algorithms". In: Proceedings of iiWAS'08, pp 591–595