An Improvised Algorithm for Relevant Content Extraction from Web Pages

Aanshi Bhardwaj
Information Technology, Panjab University
Chandigarh-160014, India
bhardwajaanshi@gmail.com

Veenu Mangat
Information Technology, Panjab University
Chandigarh-160014, India
veenumangat@yahoo.com

***Abstract*— World Wide Web (WWW) is now a famous medium by which people all around the world can spread and gather information of all kind. However, there is large amount of irrelevant redundant and information on web pages also. Such information makes various web mining tasks web page crawling, web page classification, link based ranking and topic distillation complex. Previously, the relevant content was extracted only from textual part of web pages. But now-a-days the content on web page is not only in the text form but also as an image, video or audio. This paper proposes an improved algorithm for extracting informative content from web pages i.e. it extracts the relevant content not only as text but also as images, videos, audios, adobe flash files and online games. Experiments were conducted on different real websites show that precision and recall values of our approach is superior to the previous Word to Leaf Ratio approach.**

***Index Terms*— Web Mining, Web Content Mining, Content extraction, Document object Model, ontology generation, Tag ratio, Word to Leaf ratio.**

I. Introduction

The term Web Mining was coined by Oren Etzioni in 1996 [1]. Web mining is an application of data mining techniques which help in discovering patterns from web. It is known as the knowledge discovery from web data that contains web documents, hyperlinks between documents, and usage logs of websites etc. A web page provides a collection of facts to the users that correspond to the content of a web page. It may consist of structured records such as lists and tables, text, video or audio. On the basis of which part of the web is to be mined, web mining is divided into three areas of interest. These are Web Content Mining, Web Structure Mining and Web Usage Mining. This paper focuses on an application of Web Content Mining. Web Content Mining is ta technique of extracting useful information from the web documents.

In recent years with the rapid progress of the Internet, one can easily access different kinds of information on the World Wide Web (WWW). But it has become very inconvenient for users to obtain relevant information on the WWW as information resources on the WWW is growing rapidly and there is large amount of redundant and irrelevant information on web pages. Only the relevant and informative content on web pages is of interest to the users. Commercial web sites such as e-commerce stores, news, portal sites etc. contain redundant and irrelevant information because on user's request they generate information on web pages dynamically. A lot of redundant pages like identical pages with different URL or as mirror sites also exist. Such content is added to the increase commercial value of websites and make navigation easier for users on websites. It has been found that around 50% of the content in a webpage is generally irrelevant [2]. Mainly, advertisements, copyright statements, privacy statements, logos, table of contents, navigational panel, footers and headers come under noisy content. Table of content and navigational panel are provided to make it easier for users to navigate through web pages. Table of content and navigational panel blocks are called redundant blocks because they are present on almost every web page. Therefore, a potential method is required for detecting the main content of web page. The process of identifying main content blocks from a web page is called content extraction. The term content extraction was given by Rahman [3].

Content extraction has many applications- it makes easier for users to access the information in timely efficient manner. Redundant and irrelevant information is removed. It also increases the performance of search engines because they will not waste their time and memory in indexing and storing irrelevant content. We can say, it acts as a preprocessor of web pages for search engines. It also helps users who access the web through small screen devices because they can easily point out the relevant content. It makes several web mining tasks as web page crawling, web page classification, link based ranking, topic distillation simple. It can help in generating automatically rich site summary from blogs or articles. Key information which describes a web page generated in some of the approaches can help in classifying web pages. It can also help in developing catalogue of websites by categorizing websites on the basis of key information and

doi:10.4304/jetwi.6.2.226-230

integrating them to form catalogues. The efficiency of information retrieval and information extraction systems is also increased. Information extractors can directly extract patterns from informative content rather scanning the whole page. Users who use internet through small screen devices, through content extraction it becomes easy for them to point out the relevant content. Else it would be difficult for users to get actual information as the information is displayed on small screen. Content Extraction is also being used in application like domain ontology generation. Various authors have presented different methods for extracting relevant information from web pages [4].

II. Related Work

Use of Word to leaf ratio (WLR) [5] for content extraction is a new approach. Previously Tag ratio TR [6] as in (1) was used. TR was computed for each line. Then a threshold value selects tag ratio values as content and non -content blocks. The main problem with TR approach was that code of a web page can be indented or unintended which leads to different values for tag ratios depending upon the distribution of code. Another ratio was used called chars to node ratio CNR [7] as in (2) to overcome the problems of tag ratio. CNR is dependent on the DOM structure of web page and it is not related to the distribution of code on web page. Thus, CNR solves the problem of tag ratio. But CNR purposelessly gives importance to long words. So, a new ratio called WLR was introduced which count words instead of characters. This approach evaluates according to the relevance of text. Also instead of counting all nodes of a subtree this ratio takes into evaluation only leaf nodes because these are the only nodes that contain actual text. In WLR ratio approach first a DOM tree of web page is constructed. Then noisy nodes which contain title, script, head, meta, style, no script, link, select, #comment and the nodes which have no words and are not visible are removed. From the remaining nodes WLR is computed as in (3). The nodes which have high WLR are taken as initial node set. Relevance of nodes in initial set is computed by taking those nodes only which have higher density of text and discarding other nodes. After computing relevance best node is selected which is having the highest relevance.

$$\text{TR} = \frac{\text{number of non HTML tag characters}}{\text{number of HTML tags}} \quad (1)$$

$$\text{CNR} = \frac{\text{number of characters contained in the subtree of a node}}{\text{number of nodes in this subtree}} \quad (2)$$

$$\text{WLR} = \frac{\text{number of words in the particular node}}{\text{number of leaves in subtree of that node}} \quad (3)$$

The above algorithm is efficient only for those websites which have only text as content. Nowdays, the content is represented through other means also not only as text. It becomes more clear if a web page contains the content as image and text combination both. For example on the webpages of www. Slideshare.net, the content present on it is as an image Bt till now there is no method which discusses about the content in <img>, <video> , <embed>, <object>, <canvas> and <svg> tags. We have improvised the above algorithm to be used for those websites which have content in <img>, <video>, <embed>, <object>, <canvas> and <svg> tags. Embed and object tags are used for dispalying flash enabled content on web pages. Canvas and svg tags are in html5 are for graphics and are used on the websites which are for gaming. The plugin created by the WLR ratio approach doesnot extract informative content from the above discussed tags but the proposed plugin does it efficiently.

III. Proposed Approach

Our approach has improved the word to leaf ratio (WLR) approach so that new approach can extract not only the textual content but also the content which is displayed through images, videos or audios. Following are the steps for extraction process:

1. Construct DOM (Document Object Model) tree of web page.
2. Hide noisy nodes which contain title, script, head, meta, style, no script, link, select, #comment and the nodes which have no words and are not visible.
3. Compute word to leaf ratio (WLR) as in (4).

$$\text{WLR}(n) = \frac{\text{tw}(n)}{\text{l}(n)} \quad (4)$$

Where tw (n) = number of words in the node n
l(n) = number of leaves in the subtree of node n
Count formatted line of text as a single leaf because due to formatting, the line of text comes under different node. But in actual it's a single line so count it as one leaf.

4. Obtain initial node set which contains higher density of text. Initial node set I is defined as those nodes which satisfy the condition in (5).

$$\text{WLR}(n) \geq \sqrt{\max}_{\text{WLR}} \times \text{WLR (root)} \quad (5)$$

5. Relative position of node n is defined as in (6), (7), (8) and (9).

$$\text{R}(n) = \text{WLR}(n) \times \max \left[\text{w}(n), \sum_{n_1 \in Children(n)} \text{R}(n_1) \right] \quad (6)$$

$$\text{Where w}(n) = \begin{cases} r_{pos}(n) \times r_{WLR}(n) & \text{if } n \in \text{I} \\ 0 \text{ if } n \in \text{I} \end{cases} \quad (7)$$

$$\text{Where} \quad r_{pos}(n) = 1 - \frac{\text{id}(n) - \min_{\text{id}}}{\max_{\text{id}} - \min_{\text{id}}} \quad , \quad (8)$$

$$r_{WLR}(n) = \frac{WLR(n) - min_{WLR}}{max_{WLR} - min_{WLR}} \quad (9)$$

Where min_{id}, max_{id} are the minimum and maximum values of identifiers in I, min_{WLR}, max_{WLR} are the minimum and maximum WLR values from step 3. Relevance includes those nodes which have higher density of text. If two nodes have same R (n) then the node with lower identifier is selected.

6. Selected node has the required textual information.
7. Look for the tags <img>, <video>, <embed>, <canvas>, <object> and <svg> tags whose size is greater than 120000 i.e. (more than 300 X 400 or 400 X 300). We have obtained this value by testing on various websites flicker, tumblr, facebook, youtube, slideshare, yahoo etc. and got this mean value of height for extraction.
8. Then subsequently unhide all the parents of this discovered tag by backtracking the tree and without affecting its siblings (do not unhide the siblings).
9. Selected node has the content as image, audio, video, flash file, online game.

IV. RESULTS

Figure 1 and figure 2 compares the content extraction by the proposed plugin called Extract Content and WLR plugin called Relevant content on a webpage of www.slideshare.net.

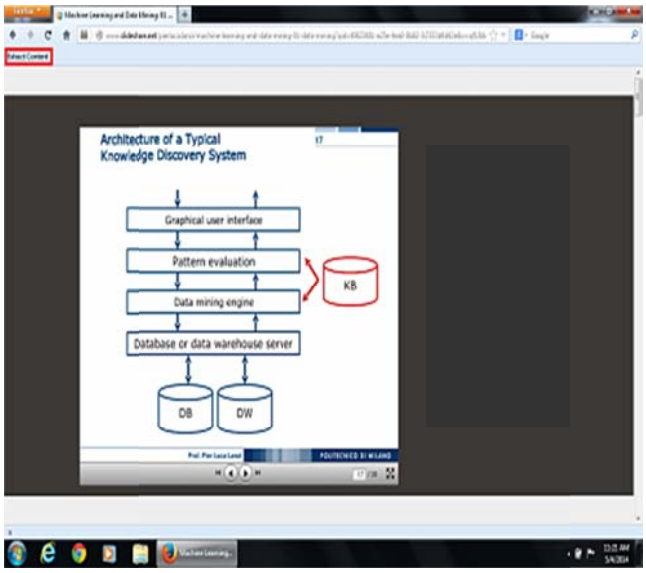

Figure1. Content extraction through Extract Content Plugin

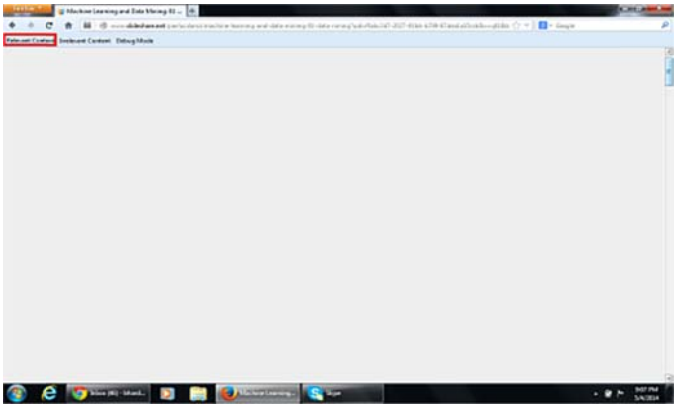

Figure2. Content extraction through Relevant Content Plugin

The metric used for the comparison is precision and recall. Precision is the fraction of retrieved blocks that are relevant to the findings.

$$\text{Precision} = \frac{\text{Relevant blocks} \cap \text{Retrieved blocks}}{\text{Retrieved blocks}}$$

Recall in information retrieval is the fraction of the blocks that are relevant to the algorithm that are successfully retrieved.

$$\text{Recall} = \frac{\text{Relevant blocks} \cap \text{Retrieved blocks}}{\text{Relevant blocks}}$$

Table 1 and 2 shows the precision and recall values from the previous algorithm and proposed algorithm respectively.

V. CONCLUSION AND FUTURE SCOPE

Automatic Content Extraction is an emerging field in research area as the amount and type of information added is increasing and changing day by day. Content Extraction is useful for the human users as they will get the required information in a time efficient manner. It helps systems like robots, indexers, crawlers, etc. in preprocessing i.e. to extract the main content of a web page and prevent the processing of noisy and irrelevant information. We have proposed a new approach for extracting the content from <img>, <video> , <embedd>, <object>, <canvas> and <svg> tags which was not earlier available. The proposed approach has shown good precision and recall values for different websites. This algorithm does not extract content from the above mentioned tags if these tags are used in the background of tags or we can say as tag's property. But this limitation is rarely found because these tags are mainly used independent of . We will try to adapt this approach to a crawler so that it can efficiently extract the refined and rich content from web pages. We also plan to incorporate hyperlink information to detect menus of a webpage.

TABLE1.
BLOCK LEVEL PRECISION AND RECALL VALUES FROM PREVIOUS ALGORITHM

URL	Relevant blocks	Retrieved blocks	Precision	Recall
https://www.youtube.com/watch?v=yGIHjTmTFfA	2	1	100%	50%
htttp://vimeo.com/channels/staffpicks/93175979	2	0	0%	0%
https://playcanvas.com	7	5	100%	71%
http://www.slideshare.net/Achievers/19-signs-of-a-disengaged-employee-by-achievers-slide-share	2	1	100%	50%
http://svg.kvalitne.cz/worldlandmarks/worldlandmarks.svg	1	0	0%	0%
https://www.flickr.com/	11	4	100%	36%
https://www.facebook.com/photo.php?fbid=10152171822119613&set=a.10150629086759613.391474.714569612&type=1	10	13	69%	90%
http://wallbase.cc/wallpaper/1937431	4	3	100%	75%

TABLE2.
BLOCK LEVEL PRECISION AND RECALL VALUES FROM PROPOSED ALGORITHM

URL	Relevant blocks	Retrieved blocks	Precision	Recall
https://www.youtube.com/watch?v=yGIHjTmTFfA	2	2	100%	100%
htttp://vimeo.com/channels/staffpicks/93175979	2	1	100%	50%
https://playcanvas.com/	7	8	88%	100%
http://www.slideshare.net/Achievers/19-signs-of-a-disengaged-employee-by-achievers-slide-share	2	2	100%	100%
http://svg.kvalitne.cz/worldlandmarks/worldlandmarks.svg	1	1	100%	100%
https://www.flickr.com/	11	9	100%	82%
https://www.facebook.com/photo.php?fbid=10152171822119613&set=a.10150629086759613.391474.714569612&type=1	10	13	77%	100%
http://wallbase.cc/wallpaper/1937431	4	4	100%	100%

ACKNOWLEDGMENT

I sincerely thank all those who helped me in completing this task.

REFERENCES

[1] O.etzioni. The world Wield Web: Quagmire or Gold Mining Communicate of the ACM, (39)11:65-68, 1996.
[2] D.Gibson, K.Punera and A.Tomkins, "The Volume and Evolution of Web Page Templates", Proceedings of WWW '05 Special interest tracks and posters of the 14th International Conference on World Wide Web, pp. 830-839, 2005.

[3] A.F.R.Rahman, H.Alam and R.Hartono, "Content Extraction from HTML Documents", International workshop on Web Document Analysis, pp. 7-10, 2001.
[4] A. Bhardwaj and V. Mangat, "A Novel Approach for Content Extraction from Web Pages", 2014 Recent Advances in Engineering and Computational Sciences, pp. 1-4, 2014.
[5] D.Insa, J.Silva and S.Tamarit, "Using the Words/leafs Ratio in the Dom tree for Content Extraction", The Journal of Logic and Algebraic Programming", vol. 82, no. 8, pp. 311-325, 2013.
[6] T. Weninger, W.H. Hsu, J. Han, CETR: Content Extraction via Tag Ratios, in: M. Rappa, P. Jones, J. Freire, S. Chakrabarti (Eds.), Proceedings of the 19th International Conference on World Wide Web (WWW10), ACM, 2010, pp. 971–980.
[7] S. Lopez, J. Silva, D. Insa, Using the DOM tree for Content Extraction, in: J. Silva, F. Tiezzi (Eds.), Proceedings of the 8th International Workshop on Automated Specification and Verification of Web Systems (WWV 12), volume 98 of EPTCS, pp. 46–59.

Aanshi Bhardwaj, pursuing Masters of Engineering in Information Technology from Panjab University. She has presented 3 papers in international and 1 paper in national conferences. Her research area includes web Mining.'

Veenu Mangat, Masters of Engineering in Computer Science, has been working as an Assistant Professor in Information Technology for the last 10 years. She is currently pursuing her Ph.D. in the area of data mining. She has 8 publications in national and international journals and has presented 5 papers in international conferences of repute. Her research areas include computational intelligence, data mining, optimization techniques and soft computing.'