A Survey on Techniques for Effectual Web Content Extraction

Sheenam Garg[1], Prof. Hiteishi Diwanji[2]
[1, 2] Department of Information Technology
[1, 2] L.D.College of Engineering, Ahmedabad, Gujarat, India

Abstract- *Internet is continuously developing and improving and has become the prime source of knowledge and information. This vast amount of data is increasing day by day. Information can be of different type like text, audio, video, etc. The web pages consist of mixture of information such as main content, copyright, navigational panel, advertisements, etc. It is very essential to differentiate important information from noisy content that may misguide users' interest. For end user only part of information which is desired is important rest all is considered as noise which degrades the information and can harm the web mining. This paper discusses various approaches for extracting informative content from web pages and removes noisy and redundant data.*

Keywords- Content Extraction, DOM Tree Generation, Repetition tags, Statistical relation, Pattern Tree, Multilevel Pages, Redundant Data, Noisy data

I. INTRODUCTION

Web Mining is a Data Mining technique to automatically discover and extract information from World Wide Web. Web Mining is used to capture relevant data about consumer, individual user and several others. The contents of Web pages are the primary focus of Web mining applications [1].Web Mining decomposed into Resource Discovery, Information Selection & Pre-processing, Generalization and Analysis. We can classify web mining in 3 types according to its mining techniques that is web structure mining, web content mining, web usage mining.

A user is mainly interested in the original content of web page so, the process of identifying and fetching main content blocks from a web page is called content extraction. The term content extraction was found by Rahman[2]. The content extraction is very useful for pre-processing the data in many fields such as web mining, recommendation system, decision making, expert system, knowledge discovery and so on. It is also useful to special tasks such as false advertisement detection, demand forecasting, and comment extraction on product reviews [3]. The DOM Based page Segmentation is used to discard the noisy content block and extract the informative content block from Web Pages. Initially a XML or HTML Web Page is converted into DOM tree and noise is removed using DOM Based Page Segmentation which converts the page into blocks and regions. Performance of Web Content extraction is analysed based on complexity and efficiency of the method. For content extraction firstly DOM tree is generated. HTML attributes, Tag pattern generation, Subject detection, Node density, Visual information, text density etc. are used for precise content extraction and removing noisy data. In this survey paper we are discussing above techniques in detail.

II. DOM TREE GENERATION

Document Object Model (DOM) [4] is a standardized, platform-independent and language-independent interface for accessing and updating content, structure and style of any web documents. We can generate DOM tree for each HTML page where tags are internal nodes and the detailed text and images are leaf nodes. For example,

```
<HTML>
<HEAD>
<TITLE> text </TITLE>
</HEAD>
<BODY>
<P> p text</P>
<IMG SRC= "1.jpg"></IMG>
</BODY>
</HTML>
```

Dom tree for above HTML code is given below:

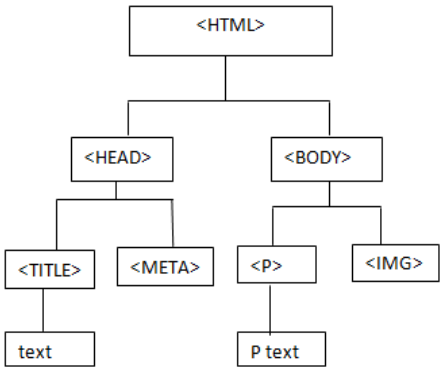

Figure 1 DOM tree

The growth of web pages on internet continues and the web Page organization is very essential. The Web Pages can be categorised into Navigation page and content page. A DOM based block text identification method proposed which detects the Navigation Page. This approach used to extracting the text segment block from a Web Page.

III. LITERATURE SURVEY

There are various techniques used for content extraction and noise removal. Each method has different percentage of content extraction and noise removal. According to the type of website i.e. government website, multilevel website, shopping website different content extraction techniques are applied for efficient and precise non-redundant content extraction and noise removal.

1. Effectual Web Content Mining using Noise Removal from Web Pages [5]

In this technique the following noises are removed step by step: (1) Primary noises such as Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and other Uninteresting Data such as audio, video, multiple links. (2) Redundant Contents and (3) Noise Contents having low block importance. These noises are removed by performing three operations. First, using the Block Splitting operation, primary noises are removed and only the useful text contents are partitioned into blocks. Second, using Simhash algorithm, the duplicate blocks are removed to obtain the distinct blocks. For each block, three parameters namely Keyword Redundancy (KR), Linkword Percentage (LP) and Titleword Relevancy (TR) are calculated. Using these three parameters block importance value (BI) is calculated. Based on a threshold value the important blocks are selected using sketching algorithm and the keywords are extracted from those important blocks.

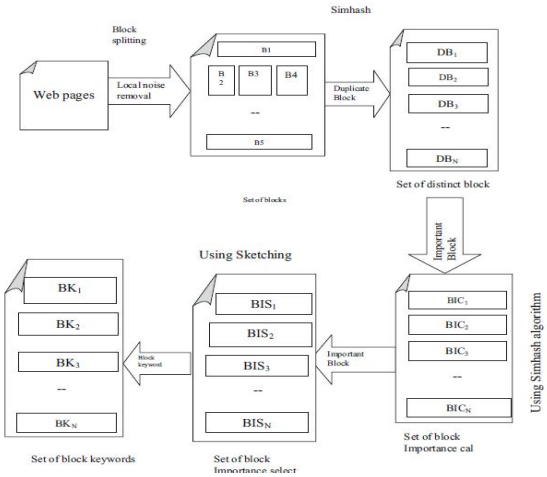

Figure 2 System Architecture

1.1 SimHash Algorithm:

```
Simhash (document D)
{
  Int vectorSim[0..(f-1)]=0;
  For (each feature F in document D) Do
      F is hashed into an f-bit hash value X;
      For (i=0;i<f;i++) Do
          If (X[t]==1) Then
             Stm[i]=Stm[i]+weight(F);
          Else
            Stm[i]=Stm[i]-weight(F);
      End
      For (t=0;t<f;t++) Do
          If ( Sim[t]>0)Then
             sim[t]=1;
          Else
             sim[t]=0;
       End
   End
}
```

1) Keywords are extracted from each block and their corresponding frequency is identified. The concept is that the most important keyword or keyword phrase will be the most frequently used keywords in a web page. With the use of this keyword and its frequency, fingerprint of each block is identified.
2) Fingerprint is generated from a block given as follows.

1. Maintain k-dimensional vector A, Initialize each dimension to zero.
2. Hash the keyword into f-bit hash value using, hashing schema.
3. f-bits (unique to the keyword) are incremented or decremented in the f components of the vector A by the frequency of that keyword as follows:

- If i-th bit of the hash value is 1, increment the i-th component of A by the frequency of that keyword.
- If i-th bit of the hash value is 0, decrement the i-th component of A by the frequency of that keyword.
- After all keywords have been processed, some components are positive while others are negative.

After calculating the fingerprints of each block, blocks having same fingerprints will be selected and only one from them will be kept.

1.2 Important Block Selection using Sketching Algorithm:

Step 1: U →space of all possible documents

Step 2: S →U: collection of documents
Step 3: Sketching: U×U → [0,1]: a similarity measure among documents.

i) If p,q are very similar Sketching(p,q) is close to 1
ii) If p,q are very dissimilar, Sketching(p,q) is close to 0
iii) Usually: Sketching (p,q) = 1-d(p,q), where d(p,q) is a normalized distance between p and q.

Step 4: t = (d-1/N) where d→ Number of distinct blocks, N→ Number of blocks
Step 5: R (Results) →R: p,q is selected if Sketching(p,q) → t (threshold)

Here the normalization technique used is Third Normal Form (3NF)

The following equation computes the block importance (bI) of each distinct block.

$$b_I = 1 - \left[\sum_{i=1}^{|n|} \frac{1}{2}K_R(i) + \frac{1}{3}L_P(i) + \frac{1}{6}T_R(i)\right], \quad where, 0 \leq b_I \geq 1$$

The representation of each parameter is as follows:

1. Keyword Redundancy,

$$K_R = \left[\frac{N/d - 1}{N - 1}\right]$$

where, N → Number of Keywords in a block
d → Number of distinct keywords in a block

2. Linkword Percentage,

$$L_P = \frac{n_l}{N}$$

where, n_l→ Number of Link Keywords in a block

3. Titleword Relevancy,

$$T_R = \frac{n_t}{\left(n_t + \sum_{i=1}^{|n_t|} F(n_t^{(i)})\right)}$$

where, n_t is the number of title keywords,
$F(n_t^{(i)})$ is frequency of the title keyword in a block

As per the calculations each parameter has assigned its own significant weightage for the best solution, which is 1/2 for KR, 1/3 for LP and 1/6 for TR.

2. Improving Web Data Extraction By Noise Removal [6]

It is generally observed that Web Pages of a single web site often follow similar layout pattern, and the noise elements are repeated in almost all web pages. In this technique, an algorithm is developed to extract the Visual Blocks of a Web page [7] of a web site using DOM and Visual Characteristics, and then it is converted to the Pattern Tree. The Pattern Tree of different web pages of a single web site is mapped to find the similarity pattern among the web pages of the website. For each node the Node Importance Measure is calculated, which is used to discriminate noise and main element of the web page.

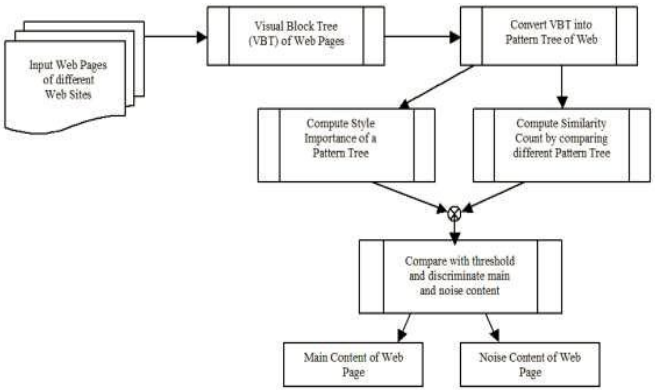

Figure 3 System Methodology adopted in design of Web Page Noise Elimination System

The noise elements of a web page is distinguished from the main content based on the following

Heuristic rules-

1) Nodes having large number of presentation styles are considered as important node and vice versa.
2) The unique pattern nodes at the specified level across web pages are referred to as important and vice versa.
3) Noise element shows consistency across the web page of a web site.

Two Node Importance metrics are constructed using the first two rule –

Style Importance Metrics: It is used to specify the number of styles applied on a given node.

Similarity Count Metrics: It is used to specify the count of node that is repeated across web pages, usually the noise element shows repeated ness, where as the main content shows uniqueness.

Using these two metrics the main content is bifurcated from the noise through the formula

$$NodeImportance(N_i) = SI_i + SC_i$$

This proposed method is depend on the heuristic that Web designer usually follow consistency while designing the web pages of the website, the noise elements like advertisement, banner, hyperlink etc are normally placed at similar position on each web page of a web site. So, this technique does not work accurately for the websites that have different patterns for some webpages.

3. Content Extraction Based on Statistic and Position Relationship Between Title and Content [7]

In previous technique consistency in the design of webpages of a website is considered. In this technique a web information extraction model based on statistical and positional relationship between the title and content is proposed. First the HTML file of the webpage is parsed and converted into DOM tree. Then each node is indexed starting from root node as zero. Next step is to calculate the attributes of each node. These attributes are: c which is content length of each node, a which is anchor length of each node, c_to_c which is ratio of single node content length to the total content length, a_to_a which is ratio of single node anchor length to the total anchor length, a_to_c which is ratio of anchor length to content length within a single node, need_ which is a Boolean type variable marking whether the node has effective texts. After the attributes are calculated the tag of <title> will be extracted as title Page. Then traverse the DOM tree from top to bottom to compare each text node with title Page.

Now, skip text node whose attribute of c is too short, filter out space and double quotation, duplicate removal. First of all, the number of how many same characters appear in both strings is calculated. Define Sim as formula (1),

$$Sim = \frac{count}{title\_len + text\_len}$$

The node which contains the right content usually possesses features like big enough c_to_c, small enough a_to_c and a_to_a. As the index of the right title, denoted as ind_t, is always greater than or equal to ind1 and less than ind2 adding up to an offset that is always five in practice. So in the range of [ind1, ind2+offset], every text is checked, and the most qualified text then will be the final right title. When ind_t is in the range of [ind2, ind2+offset], that is to say, the right title is extracted as a part of content initially.

4. Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices [8]

This approach is used for Web page segmentation by recognizing repetitive tag patterns called key patterns in the DOM tree structure of a page. Repetition-based Page Segmentation (REPS) algorithm, which detects key patterns in a page and generates virtual nodes to correctly segment nested blocks.

In REPS, Web page segmentation by using the repetition detection algorithm proceeds in 4 phases as follows.

1. Less meaningful tags such as <a>, <b>, <script>, , and “#comment” in the HTML source of the page are removed. After this pre-processing step, a Web page is represented as a DOM tree structure.
2. A sequence is taken from a DOM tree of a Web page using the tags in the child nodes of the root node. This approach of considering only one-depth child nodes and ignoring all other “deep” descendant nodes, gives the advantage of reducing computational costs while still preserving some hierarchical features of the DOM tree.
3. Generate candidate Web page blocks by using the key patterns. A key pattern is a repetitive pattern in a sequence that is longest and most frequent.
4. Finally key pattern-based Web page segmentation recognizes blocks in a page by modifying the 1-depth DOM tree into a more hierarchical structure by building *virtual nodes*.

Although this technique builds virtual nodes for nested blocks, it has difficulty in finding deeply nested blocks for some pages since the block in REPS is determined by the number of repetitions. Also, it could not control the number of blocks expected from the segmentation.

5. The Research and Implementation of Web Information Extraction Technology Based on Multi-level Pages [9]

This technique has two methods of web information extraction. The first method is width priority analysis method based on regular expressions. The second method is depth priority analysis method based on DOM tree.

a) Width priority analysis method based on regular expressions:

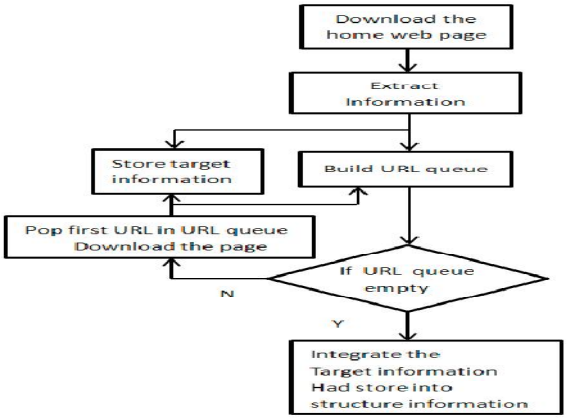

Fig.2. Width priority analysis method based on regular expressions

b) Depth priority analysis method based on DOM tree:

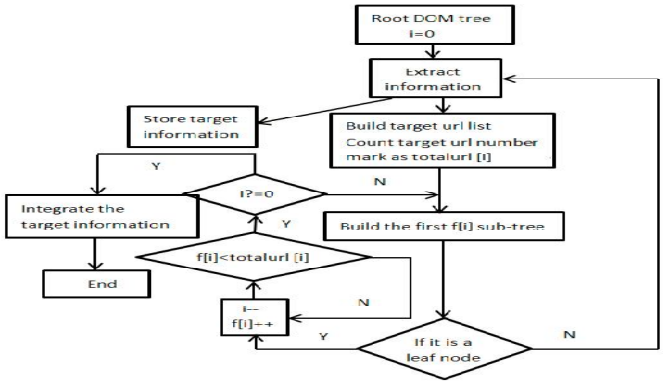

Fig.4. Depth priority analysis method based on DOM tree

The advantage of width priority analysis is that it is more flexible and drawback is that the process is too complex. The advantage of depth priority analysis is that there is no need to write a regular expression, the operation is relatively simple; Drawback is that it only apply to extract the contents of the label information for the Web and need to build a DOM tree.

Table 1 List of methods discussed

Sr No.	Technique	Method
1	Subject Detection and Node Density	Select content reach region based on subject node having maximum weight and node density using CECTD-DS.
2	Text density and visual importance of DOM nodes	Select content reach region based on maximum value of hybrid text density.
3	Word to leaf ratio with link attribute	Select content reach region based on weight and relative position of node.
4	Tag pattern generation and HTML attributes	Select content reach region based on HTML attribute value and pattern.
5	Block level elements and inline elements	Remove redundancy using density of BLE and IE blocks.

IV. CONCLUSION

In this paper, we have reviewed different techniques for informative content extraction and noise removal. Here, designing of the webpage has been given more emphasis to recognize the informative noise content. For precise and efficient content extraction and noise removal we can work on visual importance and can improve the efficiency of the DOM tree algorithm.

REFERENCES

[1] Shuang Lin, Jie Chen, Zhendong Niu, "Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction" ,TSINGHUA SCIENCE AND TECHNOLOGY, ISSN 1007-0214 05/18 pp256-264 Volume 17, Number 3, June 2012

[2] A.F.R.Rahman, H.Alam and R.Hartono, "Content extraction from HTML documents", International workshop on Web Document Analysis, pp. 7-10, 2001.

[3] Warid Petprasit and Saichon Jaiyen, "Web Content Extraction Based on Subject Detection and Node Density", 978-1-4799-6049-1/15/$31.00 ©2015 IEEE

[4] W3C Document Object Model (2009) Website. http://www.w3.org/DOM

[5] P. Sivakumar, "Effectual Web Content Mining using Noise Removal from Web Pages," Springer Science+Business Media New York 2015

[6] Neetu Narwal, "IMPROVING WEB DATA EXTRACTION BY NOISE REMOVAL", IEEE 2013

[7] Mingdong Li, Pingping Xu, Chencheng Yang "Content Extraction Based on Statistic and Position Relationship

Between Title and Content", IEEE/CIC ICCC 2014 Symposium on Social Networks and Big Data

[8] Jinbeom Kang, Jaeyoung Yang, Nonmember and Joongmin Choi, Member, "Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010

[9] Hengyu Lai, Yifei Wei, Yali Wang, Mei Song, Xiaojun Wang, "The Research and Implementation of Web Information Extraction Technology Based on Multi-level Pages", ISSC 2014 / CIICT 2014, Limerick, June 26-27