Reappearance Layout based Web Page Segmentation for Small Screen Devices

V.Kalaivani
M.E.,Ph.D.,
Associate Professor (SG),
Department of CSE,
National Engineering College,
Kovilpatti, Tuticorin(DT).

K.Rajkumar
M.E.,
Assistant Professor,
Department of CSE,
National Engineering College,
Kovilpatti, Tuticorin(DT).

ABSTRACT

Normally web sites are designed for large screen devices and hence it is not easy to browse these pages with limited user interface devices such as palm and mobile. Web page segmentation is an important technology for both search engine and web browser on mobile device. Web page segmentation is a task that breaks down the structure of web page into logical blocks which is an important step for identifying informative blocks for efficient information extraction and convenient display on the devices with small size screens. Previous repetition based segmentation method is not suitable for segmenting blocks, when there is no reappearance tags in the web pages.

In order to improve the segmentation accuracy, a new method of Reappearance Layout based Web page Segmentation (RLSE) is introduced which segments web pages based on either reappearance based segmentation (RSE) scheme or based on Layout based segmentation (LSE) such as table, div, span and frame tags , depends up on key pattern detected in the web page. If it contains reappearance tag in key pattern means, it will segment based on reappearance based segmentation. Otherwise it will segment based on web layout information. From that segmented block hyperlink is displayed on the mobile first and then user will select hyperlinks of segmented blocks based on his area of interest. The interested information alone is displayed to the user.

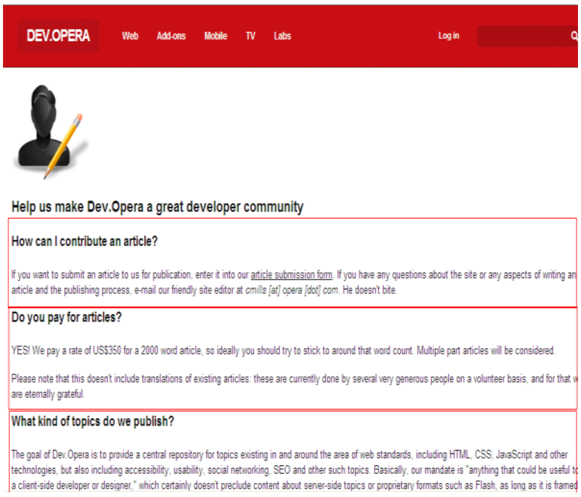
Keywords

DOM (Document object Model), Layout based segmentation (LSE), Reappearance based segmentation (RSE), RLSE (Reappearance Layout based Segmentation).

.

1. INTRODUCTION

Recently, the demand is high for smart phones portable hand-held computer and personal digital assistants. Inspite of that, mobile device has a small-display, lesser data rates, and some constraints owing to equipment limitations exist. Mobile access to web-pages with small handheld computers such as cellular phones and PDAs has been at the center of developing issues in mobile environment since early twenty first century. The web pages are developed to be displayed on a Large size screens and it is therefore not easy to display an whole web page on the small screen devices. Web page segmentation is an important technology for web driven applications such as search engine and web browser on mobile device. Web page segmentation is a technique that segments a Web page into logical blocks such that each block contains distinctive content information. Figure 1, shows an example of page segmentation where the segmented blocks are marked with boxes. This technique enhancing the performance of information extraction by ignoring noise blocks and focusing only on informative content blocks, and also facilitating the display of useful information on mobile devices with small-sized screens. Many websites do not follow standardized format, but develop Web pages.



Help us make Dev.Opera a great developer community

How can I contribute an article?

If you want to submit an article to us for publication, enter it into our article submission form. If you have any questions about the site or any aspects of writing an article and the publishing process, e-mail our friendly site editor at *cmills [at] opera [dot] com*. He doesn't bite.

Do you pay for articles?

YES! We pay a rate of US$350 for a 2000 word article, so ideally you should try to stick to around that word count. Multiple part articles will be considered.

Please note that this doesn't include translations of existing articles: these are currently done by several very generous people on a volunteer basis, and for that w are eternally grateful.

What kind of topics do we publish?

The goal of Dev.Opera is to provide a central repository for topics existing in and around the area of web standards, including HTML, CSS, JavaScript and other technologies, but also including accessibility, usability, social networking, SEO and other such topics. Basically, our mandate is "anything that could be useful t a client-side developer or designer," which certainly doesn't preclude content about server-side topics or proprietary formats such as Flash, as long as it is framed

Fig 1: Example of Web page segmentation

Early researches on Web page segmentation are based on the structure of a Web page. This template-based method is focused on measuring the structural similarities among the DOM trees of Web pages. A pitfall of this method is that it is impossible to segment information contained in a tree node because the template-based method recognizes only the structural information, not the content, of DOM tree nodes. It also needs qualitative and quantitative training examples to cover real-world situations. To overcome these difficulties,another approach was proposed by using the visual information in a Web page. This vision-based segmentation method exploits visual clues such as font size, font color,background color, spaces between paragraphs, etc.

Inspite of some successes, vision-based method heavily depends on heuristic rules that make its difficult to cope with a dynamic Web environment where the information of structural and visual are often changing. To resolve these problems, this paper proposes a Reappearance Layout based Web page segmentation (RLSE) that segments web pages based on either reappearance based scheme, by recognizing reappearance tag patterns from the DOM tree structure of a web page or web layout information.

2. RELATED WORK

Yonghyun Hwang et al [14] proposed two general solutions to this disparity exist, manual and automatic reauthoring. In manual reauthoring, Web authors prepare multiple versions of a Web page targeted to resource profiles of various platforms, including the Wireless Application Platform. Although this approach can produce high-quality pages for specific devices, it assumes a Web author will both be available to reauthor the pages.

Xing Xie et al [13] told that block importance is divided into three levels. Level 3 contains most prominent part of a page, such as headlines, main content etc. Level 2 contains useful information, but not very relevant to the topic of a page, such as navigation, directory, etc. Level 1 contains Noisy information such as ads, copyright, decoration, etc. A web page is first partitioned into semantic blocks and then spatial features, content features and that are extracted to construct a feature vector for each block.

Sandip debnath et al [9] formally defined Web page blocks and devise a new algorithm to partition an HTML page into constituent Web page blocks. They proposed four new algorithms, such as Content Extractor, Feature Extractor, K-Feature Extractor, and L-Extractor.

Peifeng Xiang et al [7] introduced an effective automatic segmentation method which combining pattern analysis and visual separators. The basic idea is that a page's semantic structure is largely reflected by repeated continuous patterns and visual separators that coincide with human's visual perception.

The template-based method builds a template with extracting rules based on regular expressions [1],[2],[4]. It collects Web pages from a target site and generates regular expression rules in order to extract content blocks through analyzing the common zone. Web pages are decomposed into blocks according to the generated rules. Whereas it is simple and easy to use with fewer errors, the template-based method requires many sample pages in the same domain to build a template. Also, it has to maintain a template for each domain. since Web pages may not be correctly processed with templates for different domain sites.

The tag-based method predefines content tags that contain useful information and finds content blocks by measuring the distance between these tags [5],[3],[13].

Lin et al [6] illustrated that the <TABLE> tag is widely used to make the structure of a Web page, and proposed a method primarily using the <TABLE> tag to extract blocks from a Web page. However, this method cannot be applied to those Web pages without the <TABLE> tags.

Yunpeng Xiao et al [15], determined the structure of web page by analyzing the page layout and then use a page-splitting algorithm to effectively partition the page into a series of tailored content blocks.

Suhit Gupta et al [10], proposed a DOM-based content extraction technique which is used to extract content from web pages which is based on exploring the DOM Tree representation of web page. However, this method cannot be applied to those Web pages without the <TABLE> tags.

WooKey Lee et al [11], proposed a new web-page search mechanism suitable for mobile devices, called an adaptive hypermedia search. It utilizes hierarchical structure of hypermedia objects for handheld devices, such as cellular phones and PDAs, which have usually limited resources.

3. PROPOSED METHOD

The proposed algorithm, named RLSE (Reappearance Layout based Web page Segmentation), detects key patterns in a Web page. If it contains reappearance tag in key pattern means, it will segment based on reappearance based segmentation. Otherwise it will segment based on web layout information such as <TABLE>, <DIV>,<SPAN> and <FRAME> tags.

The architecture of RLSE Method is shown in Figure 2. RLSE proceeds in four phases. First, a Web page is represented by a DOM tree structure. In the second phase, RLSE generates a key pattern from the DOM tree structure. In the third phase, check for key pattern. If the Key pattern contains more number of reappearance tags, then it will generate implicit node to segment nested blocks. Otherwise, it will segment based on web layout information. In the fourth phase, the hyperlinks of segmented blocks are first displayed to mobile device and then user select hyperlink based on his area of interest.

3.1 Tag Extraction

The web pages is taken from any website and stored in a folder. Then select a web page for segmentation using browse button. The content of the web page is displayed to the user. The tags are extracted from the web page. Less meaningful tags such as <a>, <b>, <script>, etc are removed.

3.2 DOM tree Construction

DOM or Document Object Model is a tree structure which represents a nesting of elements within the page and it is often used for traversing the web page. In DOM, an element is referred as a node.A DOM tree is constructed, after removing less meaningful tags such as <a>,<b>,<script>,etc

Algorithm 1:

Reappearance Layout based Webpage Segmentation (RLSE) algorithm, detects key patterns in a Web page. If it contains similar reappearance key pattern means, it will segment based on reappearance based segmentation (RSE). Otherwise it will segment based on web layout information (i.e.LSE) such as <TABLE> ,<DIV>, <SPAN> and <FRAME> tags.

RLSE algorithm

```
function RLSE (websites)
// Input: websites.
// Output: Segmented blocks.
Extract tags from the html source of web page.
function RSE (Tag list);
    if all key patterns are similar then
         //Segment the page with Algorithm A
        function pattern matching(list,reap);
    else
        //Segment the page with Algorithm B
        function LSE(Tag list);
    end if.
```

3.3 Pattern Identification

Reappearance pattern identification technique is used to find a pattern in a list or sequence that consists of HTML tag elements present in the web page.

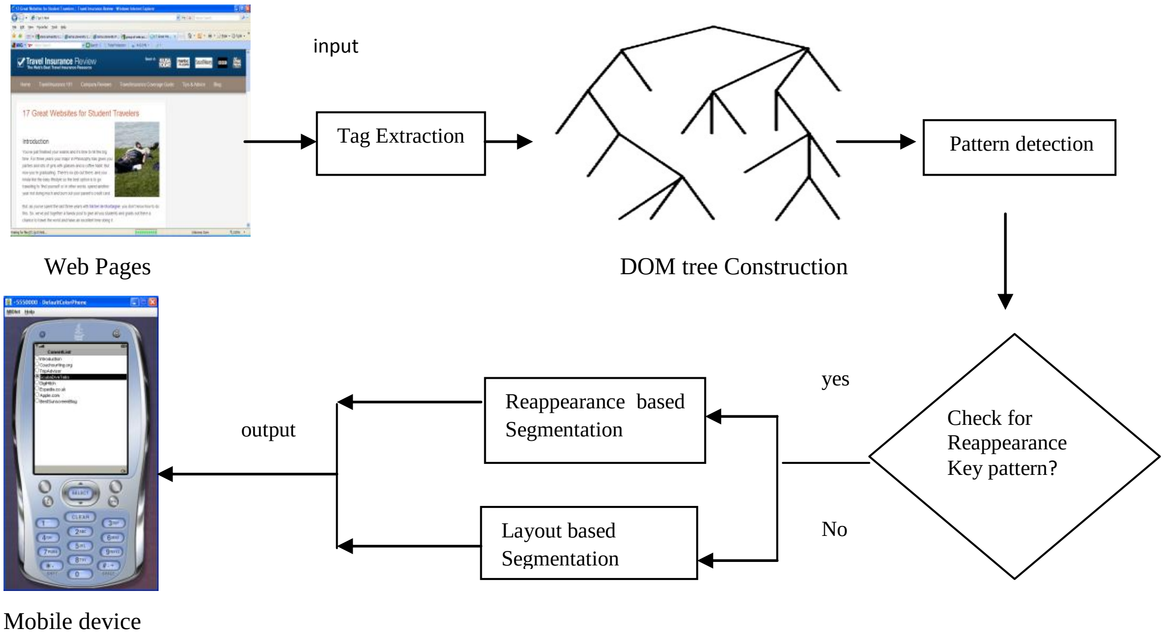

input
Tag Extraction
Pattern detection
Web Pages
DOM tree Construction
yes
Check for Reappearance Key pattern?
Reappearance based Segmentation
output
Layout based Segmentation
No
Mobile device

Fig: 2 Architecture of RLSE

Reappearance is defined as sub list of length p (>1) occurring twice or more in a list of length q. Maximum length of reappearance for a list of length is q/2, $1<p<q/2$, satisfying the formula $l<p\leq q/2$.
In Figure 3, the only repetition is cd as it occurs twice in this series. Consider this repeated subseries as a pattern. The algorithm for finding key pattern is shown below. A series whose elements are all the same must be handled exceptionally

series	Subseries (2)	Subseries(3)
c d e c d f	**c d** d e e c **c d** d f	c d e d e c e c d c d f

Fig 3: Finding reappearance pattern from a series

According to the above definition, for the series bbbbb, there could be four reappearance of bb as shown in case 1 of Figure 4. In order to be a useful pattern in a Web page, reappearance must not be overlapped with other patterns in a series. From this observation, the number of repetitions of bb in this situation should be 2 as shown in case 2 of Figure 4. After key pattern detection, it will check for pattern. If it contains more number of reappearance tag, then it will generate implicit node to segment nested blocks. Otherwise, it will segment based on web layout information.

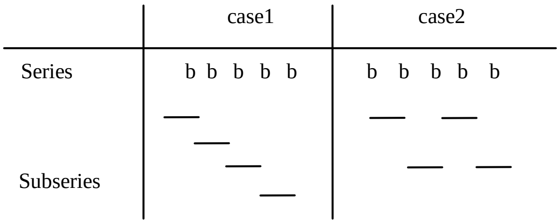


Fig 4: Finding reappearance pattern when all elements are same

For example, from the main page of WordNet , we can get a series of child nodes of the <div> node, which is “strong p p a p p strong p p p p p p div” as shown in Figure 5.

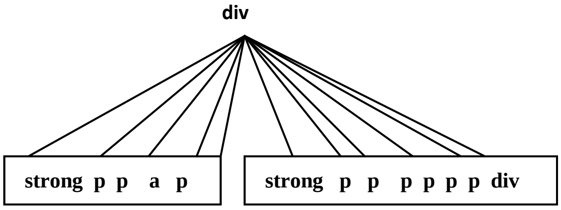


Fig 5: DOM tree structure of wordnet web page.

Algorithm 2 describes a pseudo-code for reappearance pattern detection for an input sequence *list* of length ***N***. Since the maximum length of a possible pattern is ***N****/2*, we only examine the first half of the input series by setting *maxSize* appropriately in line 5.

The sublist*(list,j)* function (line 12) returns a subseries of the input series generated from *j*th to last element of *list*. The partition*(subList,i)* function (line 13) returns all subseries of *subList* of length *i*.

Algorithm 2 Reappearance pattern Detection

```
1:  function RSE(Taglist)
2:  // Input: Tags.
3:  // Output: reappearance key pattern.
4:  $rep \leftarrow \{\ \}$
5:  $maxSize \leftarrow \frac{|N|}{2} + \begin{cases} 1 & |N|\%2 > 0 \\ 2 & otherwise \end{cases}$
6:  $i \leftarrow maxSize$
7:  while $i \geq 2$ do
8:      $candi \leftarrow \{\ \}$
9:      $AllEqual \leftarrow \{\ \}$
10:     $j \leftarrow 1$
11:  while j $\geq$ max do
12:     $subList \leftarrow$ sublist(list, j)
13:     $parts \leftarrow$ partition(subList, i)
14:    $Equal \leftarrow \{\ \}$
15:  for all series $\in$ parts do
16:    if all elements of series are same then
17:        $Equal \leftarrow Equal \cup \{series\}$
18:    else
19:       $candi \leftarrow candi \cup \{seq\}$
20:    end if
21:  end for
22:   update(AllEqual,Equal)
23: end while
24: $candi \leftarrow candi \cup maxAllEqual$
25: $reap \leftarrow reap \cup$ findCandidate(candi)
26: end while
27: return reap
28: end function
```

Note that the total number of subseries generated by partition is *N/i*. The update*(maxAllEqual,allEqual)* function (line 22) assigns the subsequence with the maximum length in the current step to *maxAllEqual* when all elements are the same. The findCandidate*(candi)* function (line 25) identifies all possible repetitions occurring at least twice in the candidate set *candi*.

3.4 Reappearance based Segmentation

A pattern matching algorithm is used for segmentation. Key patterns are used to build implicit nodes by separating the subseries containing key patterns. For doing this, a key pattern is matched with the given series from left to right sequentially.

When a match is found at position *i*, the process continues to find the next match in the remaining part of the series. If there is another match at position *j*, the subseries ranging from *i* through *j*-1 is grouped with a new implicit node. For each subseries, we add a implicit node and build a subtree with the implicit node as its root and the elements of the subseries as its children.

Algorithm A pattern Matching

```
function pattern matching (list, reap)
// Input: reappearance key pattern list.
// Output: Segmented blocks.
$blocks \leftarrow \{\ \}$
$allEqualblocks \leftarrow \{\ \}$
for all keypattern $\in$ rep do
 Series ← []
 index ← 0
 flag ← false
 allEqual ← isAllEqual(keypattern)
 i←0
 while i < $|list|$ do
    if index $\geq$ | keypattern | - 1 then
      index ← 0
      flag ← true
   end if
   if list[i] = keypattern[index] then
      if flag = true & allEqual = true then
         flag ← false
        blocks ← blocks $\cup$ {Series}
     end if
        append(Series, list[i])
        index ← index + 1
else
     if flag = true & allEqual = true then
         update(allEqualblocks, Series)
         index ← 0
    else if flag= true & allEqual $\neq$ true then
          append(Series, list[i])
     else
          index ← 0
          Series ← []
      end if
        end if
          end while
             if | Series | > 0 then
                 append(blocks, Series)
            end if
        end for
             append(blocks, allEqualSeq)
   return blocks
end function
```

Finally, from two key patterns, we can make three implicit nodes as shown in Figure 6.

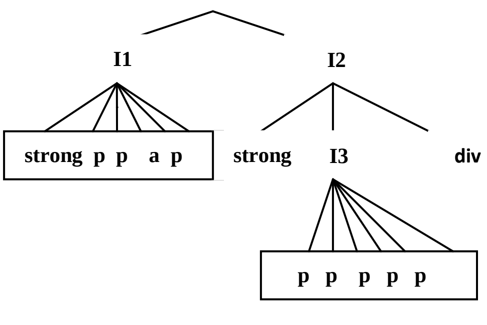


Fig 6: Generating Implicit nodes by using keypatterns.

3.5 Layout Based Segmentation (LSE)

In Layout based scheme, it will segment based on web layout information such as <TABLE>, <DIV> and <FRAME> tags. The LSE (Layout based segmentation) analyzes the <TABLE>, <TR>, <TD>, <SPAN> and <DIV> tags, which are not only used to compose the table, but also to define the layout of the Web page. After identifying the tags such as <TABLE>, <DIV>,<FRAME> ,it can be easily segmented based on the layout.

Algorithm B LSE algorithm

```
function LSE (Tag list)
// Input: Tag list [<li>,<div>,<span>-----------]
set len as length of Tag list array.
i ← 0
while i < len
j ←0
if a[i] is equal to <div> then
  Repeat
    a[j++]=a[i]
i ←i+1;
until a[i] != </div>
else if a[i] is equal to <li> then
  Repeat
    a[j++]=a[i]
i ← i+1;
until a[i] != </li>
else if a[i] is equal to <span> then
  Repeat
    a[j++]=a[i]
i ← i+1;
until a[i] != </span>
else if a[i] is equal to <table> then
  Repeat
    a[j++]=a[i]
i ←i+1;
until a[i] != </table>
else if a[i] is equal to <ul> then
  Repeat
    a[j++]=a[i]
i ← i+1;
until a[i] != </ul>
// Here a[i] is considered as layout for segmenting blocks.
end if
end while
end function
```

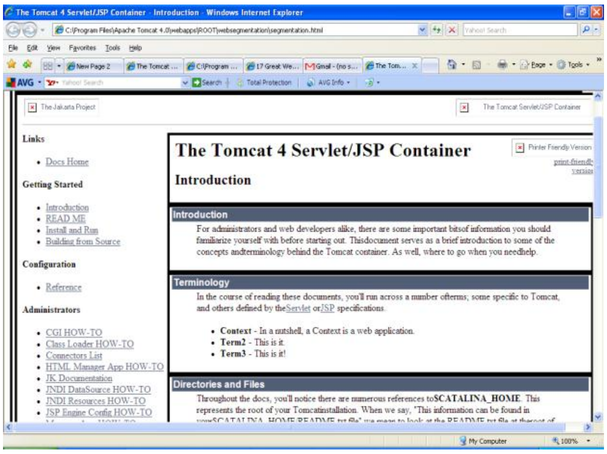
The problem of reappearance based segmentation is solved. In reappearance based segmentation scheme, it will segment more n.o of blocks, when there is less number of reappearance tags in the web page.

3.6 Display hyperlink in mobile device

After segmentation process, the web page is divided into number of blocks. Informative blocks are determined by evaluating the amount of information in the block, which can be done by assigning an importance weight to each node considering the number of reappearance of node patterns in a Web page. For simple calculation, normalize the importance weight for each node by the maximum number of repetitions. From that informative segmented block, hyperlink is created and displayed on the mobile device first. User selects the hyperlink based on his area of interest. The interested area alone is displayed to the user.

4. EXPERIMENTAL RESULTS

The performance of our RLSE algorithm for Web page segmentation is evaluated with four different types of Web sites; article, kids, portal, and shopping mall-style as shown in Figure 7.


The Tomcat 4 Servlet/JSP Container - Introduction - Windows Internet Explorer
C:\Program Files\Apache Tomcat 4.0\webapps\ROOT\websegmentation\segmentation.html
Yahoo! Search
File Edit View Favorites Tools Help
New Page 2
The Tomcat ...
C:\Program ...
17 Great We...
Gmail - (no s...
The Tom...
Page
Tools
AVG
Search
Total Protection
AVG Info
The Jakarta Project
The Tomcat Servlet/JSP Container
Printer Friendly Version
print-friendly version
My Computer
100%

Links

- Docs Home

Getting Started

- Introduction
- READ ME
- Install and Run
- Building from Source

Configuration

- Reference

Administrators

- CGI HOW-TO
- Class Loader HOW-TO
- Connectors List
- HTML Manager App HOW-TO
- JK Documentation
- JNDI DataSource HOW-TO
- JNDI Resources HOW-TO
- JSP Engine Config HOW-TO

The Tomcat 4 Servlet/JSP Container

Introduction

Introduction

For administrators and web developers alike, there are some important bitsof information you should familiarize yourself with before starting out. Thisdocument serves as a brief introduction to some of the concepts andterminology behind the Tomcat container. As well, where to go when you needhelp.

Terminology

In the course of reading these documents, you'll run across a number ofterms; some specific to Tomcat, and others defined by theServlet orJSP specifications.

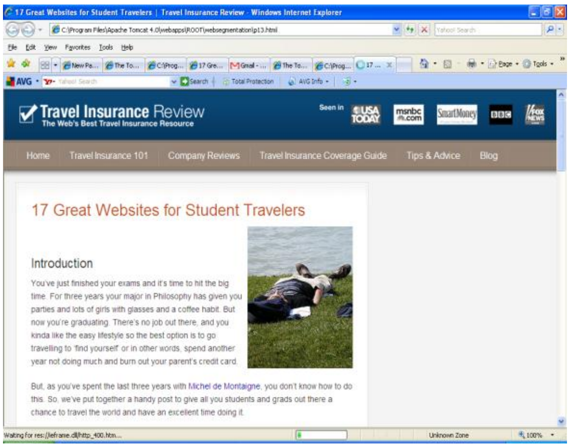
- **Context** - In a nutshell, a Context is a web application.
- **Term2** - This is it.
- **Term3** - This is it!

Directories and Files

Throughout the docs, you'll notice there are numerous references to**$CATALINA_HOME**. This represents the root of your Tomcatinstallation. When we say, "This information can be found in your$CATALINA_HOME/README.txt file" we mean to look at the README.txt file at theroot of

Fig 7: Reappearance based Segmentation
(A Tomcat web site)

The key pattern detected by pattern detection algorithm is similar, in case of Tomcat web page. Hence it is segmented based on reappearance tag present in the web page.


17 Great Websites for Student Travelers | Travel Insurance Review - Windows Internet Explorer
C:\Program Files\Apache Tomcat 4.0\webapps\ROOT\websegmentation\p13.html
Yahoo! Search
File Edit View Favorites Tools Help
New Pa...
The To...
C:\Prog...
17 Gre...
Gmail - ...
The To...
C:\Prog...
17 ...
Page
Tools
AVG
Search
Total Protection
AVG Info
Travel Insurance Review
The Web's Best Travel Insurance Resource
Seen in
USA TODAY
msnbc.com
SmartMoney
BBC
FOX NEWS
Home
Travel Insurance 101
Company Reviews
Travel Insurance Coverage Guide
Tips & Advice
Blog
Waiting for res://ieframe.dll/http_400.htm...
Unknown Zone
100%

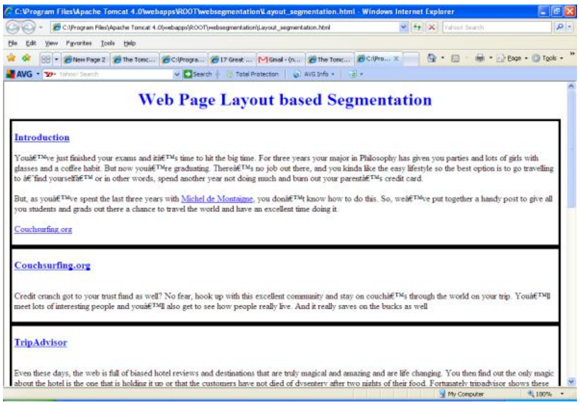
17 Great Websites for Student Travelers

Introduction

You've just finished your exams and it's time to hit the big time. For three years your major in Philosophy has given you parties and lots of girls with glasses and a coffee habit. But now you're graduating. There's no job out there, and you kinda like the easy lifestyle so the best option is to go travelling to 'find yourself' or in other words, spend another year not doing much and burn out your parent's credit card.

But, as you've spent the last three years with Michel de Montaigne, you don't know how to do this. So, we've put together a handy post to give all you students and grads out there a chance to travel the world and have an excellent time doing it.

Fig 8: Before Segmentation:
(A Travel insurance Review website)


C:\Program Files\Apache Tomcat 4.0\webapps\ROOT\websegmentation\Layout_segmentation.html - Windows Internet Explorer
C:\Program Files\Apache Tomcat 4.0\webapps\ROOT\websegmentation\Layout_segmentation.html
Yahoo! Search
File Edit View Favorites Tools Help
New Page 2
The Tomc...
C:\Progra...
17 Great ...
Gmail - (n...
The Tomc...
C:\Pro...
Page
Tools
AVG
Search
Total Protection
AVG Info
My Computer
100%

Web Page Layout based Segmentation

Introduction

Youâ€™ve just finished your exams and itâ€™s time to hit the big time. For three years your major in Philosophy has given you parties and lots of girls with glasses and a coffee habit. But now youâ€™re graduating. Thereâ€™s no job out there, and you kinda like the easy lifestyle so the best option is to go travelling to â€˜find yourselfâ€™ or in other words, spend another year not doing much and burn out your parentâ€™s credit card.

But, as youâ€™ve spent the last three years with Michel de Montaigne, you donâ€™t know how to do this. So, weâ€™ve put together a handy post to give all you students and grads out there a chance to travel the world and have an excellent time doing it.

Couchsurfing.org

Couchsurfing.org

Credit crunch got to your trust fund as well? No fear, hook up with this excellent community and stay on couchâ€™s through the world on your trip. Youâ€™ll meet lots of interesting people and youâ€™ll also get to see how people really live. And it really saves on the bucks as well

TripAdvisor

Even these days, the web is full of biased hotel reviews and destinations that are truly magical and amazing and are life changing. You then find out the only magic about the hotel is the one that is holding it up or that the customers have not died of dysentery after two nights of their food. Fortunately tripadvisor shows these

Fig 9:After Segmentation

The key pattern detected by pattern detection algorithm is not similar, in case of Travel insurance review web site. Hence it is segmented based on Layout tag present in the web page.To validate the efficiency of the proposed method, various experiments were conducted to evaluate the segmentation accuracy, its usability, and the performance on small screen devices. In order to estimate segmentation accuracy, the results of the conventional and proposed methods were compared by conducting segmentation experiments on popular Web sites.

The First step is to evaluate the segmentation accuracy of the proposed method. From that, the results are compared with conventional method. After this, the conventional reappearance based method and the proposed Reappearance Layout based Web page Segmentation method are referred to as "*RSE*" (reappearance based segmentation) and "RL*SE*" (Reappearance Layout based Web page Segmentation), respectively. From the 10 numbers of web sites, 50 web pages were collected for evaluation.

Table 1: Websites Used for Evaluation

S.No	Websites
1.	http://dev.opera.com
2.	http://www.merges.net
3.	http://www.web-datamining.net/
4.	http://www.bestindiansites.com/online-flight-booking/
5.	http://www.shoppingsites.com/
6.	http://www.kidsites.com
7.	http://www.travelinsurancereview.net/
8.	http://tomcat.apache.org/tomcat-4.1-doc/index.html
9.	http://www.merges.net/theory.html
10.	http://www.kidsites.com/sites-du/animals.htm

Evaluation can be done, with the help of two parameters namely precision and recall. The web sites used for evaluation are mentioned above in Table 1. The comparison results are analyzed and described in Table 2.

$$\text{Precision(p)}= \frac{(a)\ Number\ of\ "Correct"\ Segments}{(c)\ Number\ of\ All\ "Correct"\ Segments} \qquad (4.1)$$

$$\text{Recall (R)} = \frac{(a)\ Number\ of\ "Correct"\ Segments}{(b)\ Number\ of\ "All"\ Segments} \qquad (4.2)$$

(a) Number of "**Correct**" Segments
The number of positions segmented correctly.
(b) Number of **All** Segments.
The number of all positions segmented inspite of correct or not.
(c) Number of All "**Correc**t" Segments
The number of all positions segmented correctly.

In order to measure the whole accuracy, F-measure is used. The F-Measure can be calculated as

$$\text{F-Measure(F-M)}= \frac{2 *\ \Pr ecision * \operatorname{Re} call}{(precision + \operatorname{Re} call)} \qquad (4.3)$$

Table 2: Comparison of Proposed Methodology

Websites	REPS			RLSE		
	P	R	F-M	P	R	F-M
Travel insurance review	0.5	0.66	0.56	**1**	**1**	**1**
Dev.opera.com	0.63	0.83	0.72	**0.70**	**0.89**	**0.78**
shoppingsites .com	0.75	0.83	0.78	**0.88**	**0.85**	**0.86**
www.kidsites .com	0.88	1	0.93	**0.98**	**1**	**0.98**
Mean	0.69	0.83	0.75	**0.89**	**1**	**0.90**

The precision values of four different web sites were calculated for two methods such as Reappearance based segmentation and Reappearance Layout based Web page Segmentation and it is compared. RLSE method had better precision values than RSE. The segmentation accuracy is increased and it's shown in Figure 10.

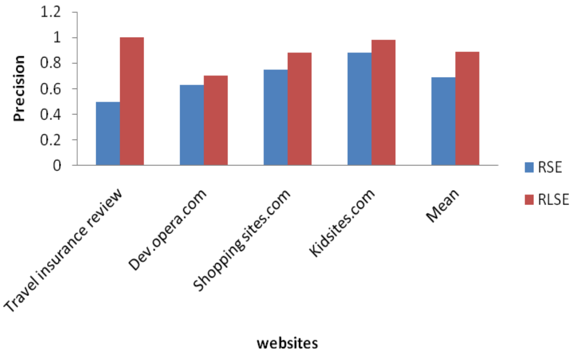


Fig 10: Comparison of RSE and RLSE based on Precision

Similarly the Recall values are also calculated for both Reappearance based segmentation method (RSE) and Dynamic web page segmentation. The segmentation accuracy is increased and it's shown in figure 11.

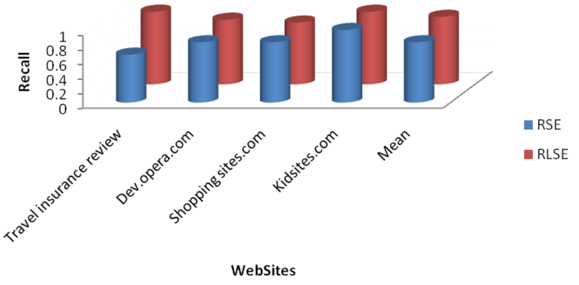


Fig 11: Comparison of RSE and RLSE based on Recall

The precision and recall values of DWS are high than RSE. Hence our method outperforms than RSE. Even though in order to measure the whole segmentation accuracy, F-Measure is calculated based on the equation (4.3)

To measure the whole segmentation accuracy, F-Measure is calculated based on the equation (4.3). The result is analyzed and compared with REPS and DWS. DWS achieves better segmentation accuracy than REPS and it is shown in figure 12

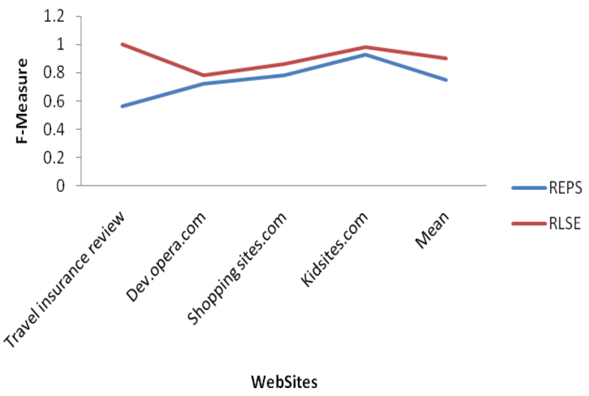
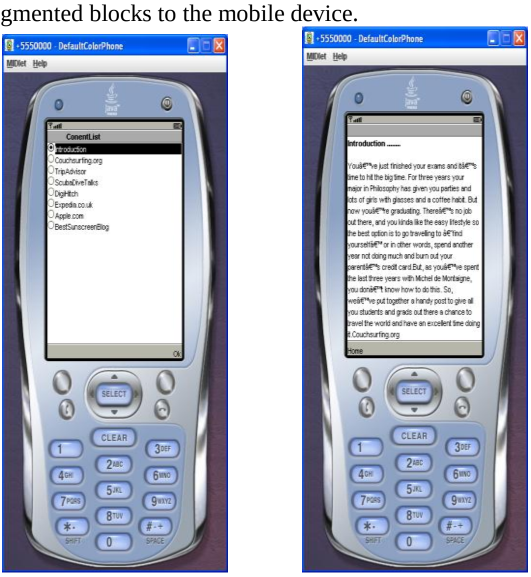


Fig 12: Comparison of RSE and RLSE based on F-Measure

After segmentation process, the web page is divided into number of blocks. Informative blocks are determined by evaluating the amount of information in the block, which can be done by assigning an importance weight to each node considering the number of reappearance of node patterns in a Web page. The bandwidth is saved, when we display the segmented blocks to the mobile device.



ConentList

● Introduction
○ Couchsurfing.org
○ TripAdvisor
○ ScubaDiveTalks
○ DigiHitch
○ Expedia.co.uk
○ Apple.com
○ BestSunscreenBlog

Introduction ....

Youâ€™ve just finished your exams and itâ€™s time to hit the big time. For three years your major in Philosophy has given you parties and lots of girls with glasses and a coffee habit. But now youâ€™re graduating. Thereâ€™s no job out there, and you kinda like the easy lifestyle so the best option is to go travelling to â€˜find yourselfâ€™ or in other words, spend another year not doing much and burn out your parentâ€™s credit card.But, as youâ€™ve spent the last three years with Michel de Montaigne, you donâ€™t know how to do this. So, weâ€™ve put together a handy post to give all you students and grads out there a chance to travel the world and have an excellent time doing it.Couchsurfing.org

Home

13 (i) Content Items 13 (ii) The first content block in (Travel insurance review website) **Fig 13: Extracted blocks as accesed on Mobile device**

For example, a web page containing a total volume of 250KB.When user wants to see this page on small screen devices, without using our system as interface. At first whole page is loaded, user had some difficulties in finding their area of interest.

When user accessed the same page with our system (RLSE) as an interface he receives only list of content blocks as shown in Figure 13 (i). Getting such content blocks will results in lower KB of bandwidth such as 20KB. Suppose that user interested only in seeing first block alone in the web page means, he selects that category from content blocks and it is alone displayed on mobile as Shown in Figure 13 (ii), for each content block, there must be around 20KB.

Table 3: Consumption of Bandwidth

Websites	Without Interface	With interface	savings
Travel Insurance.com	28 KB	4KB	**85.7%**
Shopping sites.com	32KB	4KB	**87.5%**
BestIndianSites.com	250 KB	20KB	**92%**

Over entire series of operations, the total data downloaded around just 20KB, which results in 92% savings in bandwidth. Because advertisements are also removed from the web page. User interested are alone is displayed on mobile device. Table 3 above shows the consumption of band width results for segmented blocks in a web page.

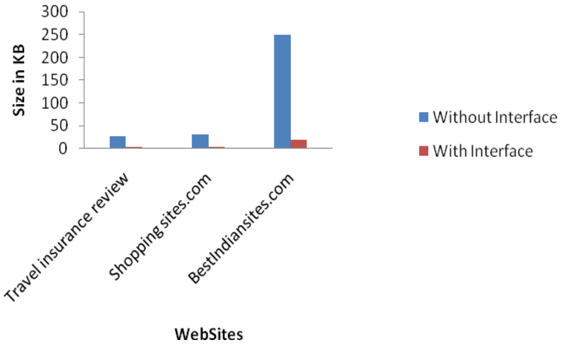


Fig 14: Bandwidth utilization

Figure 14 shows the comparison of bandwidth utilization using with interface and without interface. Using our system (RLSE) method as an interface (or) middleware, the consumption of bandwidth is reduced.

5. CONCLUSION AND FUTURE WORK

This paper proposed Reappearance Layout based web page segmentation (RLSE) for mobile device is introduced. Previous methods for Web page segmentation are not flexible, when there is less number of reappearance tags present in the web page. DWS is suitable for dynamic web environment either by recognizing reappearance tag patterns or web Layout information. From that segmented block hyperlink is displayed on the mobile device first and then user select hyperlinks based on his area of interest. The interested information alone is displayed to the user. DWS algorithm is evaluated for 50 web sites and its gives good segmentation accuracy. Even for some personal homepages and article pages that maintain no structural layout. So it's difficult to segment those web pages. In future work, we plan to develop an algorithm which combines both techniques such as Reappearance as well as layout based method for segmenting websites that don't have structural Layout.

6. REFERENCES

[1] Arasu. A and Garcia-Molina. H , "Extracting structured data from web page," *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp.337–348, 2003.

[2] Crescenzi. V, Meriald. P, and Missier. P, “Fine-grain web site structure discovery,” *Proc. 5th ACM Intl. Workshop on Web Information andData Management*, pp. 15-22, 2003.

[3] Cheolhee Choi, Jinbeom Kang and Joongmin Choi, “Extraction of user-defined data blocks using the regularity of dynamic web pages,” *Lecture Notes in Computer Science*, vol. 4681, pp. 123-133, 2007.

[4] Guohua Hu,“Study to Eliminating Noisy Information in Web Pages based on Data Mining”, *Proc. Intl. Conf. on Natural Computation* pp.660-663,2010.

[5] Hattori. G, Hoashi. K, Matsumoto. K., and Sugaya. F, “Robust web page segmentation for mobile terminal using content distances and page layout information,” Proc. 16th *Intl. Conf. on World Wide Web*, pp. 361-370, 2007

[6] Lin, S.-H., Ho, J.-M., Discovering Informative Content Blocks from Web Documents, In Proceedings of ACM SIGKDD'02, 2002

[7] Peifeng Xiang et al, “Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens.”, Proc. *Intl. Conf.* on web intelligence,2006.

[8] Riadh May. H, Akram and Othman. M,“Web Content Adaptation system”,*International Journal of Computer A*pplication vol.23,no.9,pp. 32-39,2011.

[9] Sandip Debnath, Mitra. P, “Automatic Identification of Informative Sections of Web Pages”, *IEEE Transactions on knowledge and Data Engineering* , vol.17, no.9,pp. 1233-1246, 2005.

[10] Suhit Gupta, Gail Kaiser, David Neistadt, Peter Grimm, “DOM-based Content Extraction of HTML Documents “,Proc. 12th International Conference on World Wide Web,pp. 207-214, 2003.

[11] Wookey Lee, Sanggil Kang, Seungkil Lim, “Adaptive hierarchical surrogate for searching web with mobile devices,” *IEEE Trans. Consumer Electron,* vol. 53, no. 2, pp. 796-803,2007.

[12] Xin yang,Peifeng Xiang,“Semantic HTML Page Segmentation using Type Analysis”, *Proc. 1st Intl. Conf.* on Natural computation,2006

[13] XingXie .X , Miao .G, Song, Wen, “Efficient browsing of web search results on mobile devices based on block importance model,”*IEEE Transactions on Pervasive computing and applications,*pp. 669-674, 2006.

[14] Yonghyun Hwang etal, “Structure-Aware Web Transcoding for Mobile Device” ,IEEE Transactions on Internet computing,pp. 14-21 ,2003.

[15] Yunpeng Xiao, “A Dynamic Web Page Adaptation for Mobile Device Based on Web2.0”, *IEEE Transactions on web intelligence*,pp. 119-122 , 2008

AUTHOR’S PROFILE

Dr.V.Kalaivani received the Ph.D degree in Information and Communication Engineering at Anna University Chennai, 2011. She is an Associate professor (Senior Grade) in Department of Computer Science and Engineering at National Engineering College, Kovilpatti. Her teaching experience is 15 years and research experience is 10 years. Her Research area covers Web Mining and Data Mining. She has published 6 research papers in referred international journal and 20 papers in International and National Conference Proceedings. She is a coordinator for various national conference and seminars.

K.Rajkumar is the Assistant Professor in Department of Computer Science and Engineering at National Engineering College,Kovilpatti. His teaching experience is one year. His research area covers Web Mining and Data Mining. He has attended one national and International conference in Reputed institutions. He has also published one paper in international IEEE Conference proceedings conducted by Madras institute of Technology,Anna University,Chennai