Detecting Web Page Structure for Adaptive Viewing on Small From Factor Devices

Authors: Yu Chen, Wei-Ying Ma, Hong-Jing Zhang
Microsoft Research Asia

Paper Code: UA-2
Presenter: Thong Chanchaem

Introduction

Problems

• Most web page today have been designed for desktop computer, so it is often too large to fit into the small screen of a mobile device.

• The user have to manually scroll the window to find the content of interest and position the window properly for reading information.

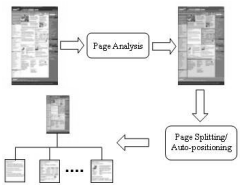
• This is a time consuming process.

Goal and Solution

Goal : to find a better way to enable easy navigation and browsing of a large web page on a small-form-factor device.

Solution:
Two level hierarchy with a thumbnail representation

Their Approach

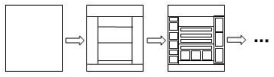


Two Process

1. Page Analysis : analyzes the structure of a give web page.
2. Page Splitting : splits the web page into a two-level hierarchy.

Page Analysis

Page Analysis : to extract the semantic structure of an existing web page.

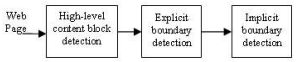


The process of identifying the content blocks.

Page Analysis :
-Content Structure Embedded by the web author.
-High-Level Content Block Detection.
  -Selecting Nodes.
  -Detection of Header and Footer.
  -Detection of left and Right Side Bar.
-Explicit Separator Detection.
-Implicit Separator Detection.

Content Structure Embedded by the web author

To detect the high-level content blocks (locations and sizes of header, footer, side bar and body) and split the content blocks .



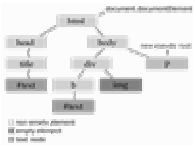
The three-step processing in their web page analysis algorithm

- Explicit separators are created using certain HTML tags.
- Implicit separators are created by leaving a blank space between content in a web page.

High-Level Content Block Detection

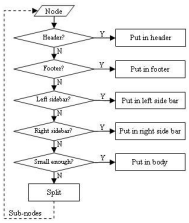
1. Selecting Node

Example of HTML DOM tree



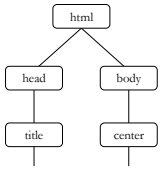
Picture from O'Reilly

Selecting appropriate nodes and classifying them into one of the five high level content blocks



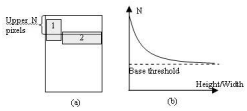
Yahoo! Homepage

```
.
.
.
adve.style.display=(adve.getElementsByTagName('a')[0].href=opts[id][2]?'inline':'none';
perfe.style.display=(perfe.getElementsByTagName('a')[0].href=opts[id][3]?'inline':'none';
}
</script><![endif]-->
</head>
<body id=bod topmargin=7 marginheight=7>
<center>
<map name=m><area alt="My Yahoo!" coords="44,0,106,47" href=r/i1>
<area alt=Finance coords="121,0,170,47" href=r/f1
.
.
.
<img src=http://us.i1.yimg.com/us.yimg.com/i/space.gif width=685 height=1>
</td></tr>
</table>
</center>
<form name=hf><input type=hidden name=h></form>
<script language=javascript>now=new Date;t4=now.getTime();</script>
</body></html>
```



High-Level Content Block Detection (cont.)

2. Detection of Header and Footer



Dynamic threshold for header and footer detection

$N = base\_threshold + F(Height/Width)$
where $F(x) = a/(b*x+c)$
and *base_threshold*, a, b and c are constants.

From their experiment, the best performance can be achieved by setting
base_threshold = 160, $a = 40$, $b = 20$, and $c=1$.

High-Level Content Block Detection (cont.)

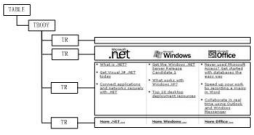
3. Detection of Left and Right Side Bar

The result of high level content block detection on the Microsoft's home page

In their experiments, they define the left 1/4 part of a web page to be the left side bar region, and right 1/4 part to be the right side bar region.

Explicit Separator Detection

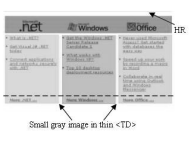
Explicit separators can be detected by analyzing the properties of the tags.



The DOM structure of the third node in the body block

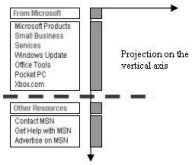
The following three types of explicit separators

1. <*HR*> represents a horizontal line
2. <*TABLE*>, <*TD*> and <*DIV*> have border properties.
3. Using an image.



Implicit Separator Detection

Implicit separators are blank areas created intentionally by the author to separate content.

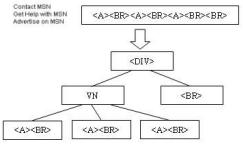


Projection on the vertical axis

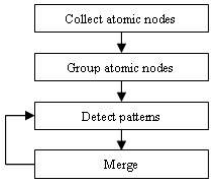
Detecting implicit separators

The first picture is a *<DIV>* containing a sequence "*<A>
<A>
<A>

". So the sequence "<A>
*" can be grouped together.



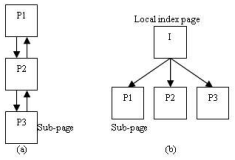
Implicit Separator Detection (cont.)



The pattern recognition algorithm is used to produce basic blocks for implicit separator detection.

Page Adaptation

There are two method for splitting a web page
- Single-subject splitting
- Multi-subject splitting



(a) Single-subject splitting breaks the whole web page into several sub-pages and connects them with next/back hyperlinks.

(b) Multi-subject splitting generates a new index page in addition to sub-pages. The index page contains hyperlinks pointing to each sub-page.

Page Adaptation (Cont.)

1. Page Splitting and Sub-page Generation

To solve the problem of style and hyperlink

- Dealing with Style
Problem : The sub-page may be lose some style information.
Solution : Copy the <header> section of a web page into each generated sub-page.
Example:

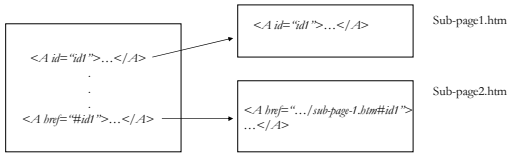
```
<head>
        <style type="text/css">
                h1 {color: red}
                h3 {color: blue}
        </style>
</head>
```

Or

```
<head>
        <link rel="stylesheet" type="text/css"
        href="/stdtheme.css" />
</head>
```

Page Adaptation (Cont.)

- Dealing with Internal Hyperlinks
Problem : The sub-page lose Internal hyperlinks.
Solution : Change the pointer in call sub-page to the proper target sub-page.
Example:



Page Adaptation (Cont.)

- Dealing with Relative Hyperlink Resolution
Problem : The sub-page lose absolute hyperlink or relative hyperlink.
Solution : Copy the <header> section into each sub-page.
Example:

Assume that the absolute address for an image is:

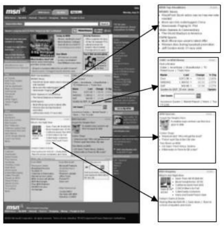
Specifies a base URL for all of the links in a page
<head>
<base href="http://www.abc.com/images/" />
</head>

The relative address

Page Adaptation (Cont.)

2. Index Page Generation

An index page which contains a thumbnail and hyperlinks to its sub-pages.

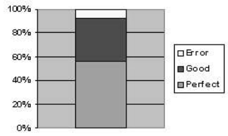


An example of the index page and sub-pages generated from the homepage of MSN.com.

Experiment and Results

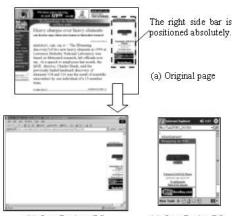
They selected 50 popular web sites and 200 typical web pages as their test data.

	Page Analysis	Page Splitting
Perfect	Perfect	Perfect
Good	Correct	Error (do not effect viewing)
Error	Error	Error



Web Site	Perfect(%)	Web Site	Perfect(%)
Yahoo.com	60%	Msnbc.com	17%
Msn.com	50%	Jobsonline.com	0%
Aol.com	60%	Flowgo.com	100%
Microsoft.com	20%	Earthlink.net	25%
Altavista Search Services	80%	Americangreetings.com	75%
Passport.com	75%	Ivillage.com	100%
Hotmail.com	60%	Mypoints.net	100%
Go.com	80%	Cnn.com	20%
Netscape.com	0%	Goto.com	0%
Amazon.com	60%	Bizrate.com	75%
Excite.com	40%	Mapquest.com	67%
Nbci.com	100%	Passthison.com	0%
Ebay.com	80%	Weather.com	75%
Bluemountain.com	80%	Mo.net	50%
Lycos.com	0%	Collsavings.com	100%
Real.com	60%	Infospace.com	100%
Looksmart.com	80%	Iwin.com	60%
Cnet.com	80%	Espn.com	0%
Angelfire.com	40%	Colonize.com	25%
Tripod.com	80%	Travelocity.com	75%
Askjeeves.com Search & Services	80%	Windowsmedia.com	0%
About.com	100%	Women.com	0%
Speedyclick.com[1]	N/A	Disney Online	33%
Iwon.com	60%	Zmedia.com	100%
Zdnet.com	80%	Google.com	50%

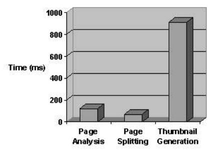
Problem cause by absolute positioning



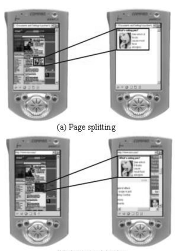
(b) On a Desktop PC (c) On a Pocket PC

Most of the splitting errors in the category of "Good" are related to styles (including CSS), absolute positioning, or scripts used to display dynamic menus.

Processing time for page adaptation algorithm



Auto-positioning instead of splitting



(a) Page splitting

(b) Auto-positioning

For a web page that uses scripts extensively, it is better not to split it.

Advantage / Disadvantage

Advantage

- Reduce client loading time.
- Reduce the consumption of network bandwidth.
- Reduce the client computation.
- Don't have to re-design web page for specific client device.
- Don't destroy the structure of source page.

Disadvantage

- Increase the computation at server side.

Critique

• The paper is well organize.
• The authors provided a lot of examples and explanations.
• Some pictures are hardly to see in detail.
• There is a lot of information in the top level that is provided a global view and index to sub-pages. Therefore, it may be difficult for user to find the content of interest in small screen devices.

Quiz

- Why can't we keep the node <CENTER> in the Yahoo! Homepage as a whole? *(see 3.2.1)*
-What are the differences between explicit and implicit separator detection?
-What is the difference between Single-subject and Multi-subject splitting?
-What is the problem that causes by CSS in "Page splitting and sub-page generation"? and How to solve it?
- When do we use auto-positioning instead of splitting?