Analysis of HTML Structure in E-commerce Websites Using Tree Representation

Jose E. Ventura* · Jeong-Sun Park*
*Dept of Industrial and Management Engineering, Myongji University

트리 표현을 사용하는 전자상거래 웹 사이트의 HTML 구조 분석

호세 벤토라* · 박 정 선*
*명지대학교 산업경영공학과

Abstract

개인화된 제품과 서비스에 대한 소비자의 요구는 성공적인 전자상거래 플랫폼을 기반으로 하고 있다. 성공적인 전자상거래 플랫폼을 개발함에 있어 자주 간과되고 있는 중요한 요소는 바로 웹 페이지의 HTML 구조이다. HTML 구조는 전자상거래 웹사이트의 속도와 랭킹을 결정짓는 기본적인 요소이다. 본 논문은 HTML 구조를 분석하기 위한 효율적이고 다소 생소한 시각화 기법을 제안하는데, 이러한 기법을 사용하여 개발자는 잠재적인 프로그래밍 오류와 개선 사항을 발견할 수 있다. 본 논문은 하나의 사례를 이용하여 제안된 기법을 더욱 구체화 시킨다.
Keywords : HTML Structure, E commerce, Tree Representation.

1. Introduction

HTML documents consist of elements which are constructed with tags that define the body of the page, paragraphs, links, tables, images, forms and others. The HTML hierarchy (DOM: Document Object Model) is a parent child sibling relationship.

An element that is directly above another element in the hierarchy is called parent of the element below it. The element below the parent is called the child. When two elements are equal in the hierarchy, they are known as siblings. The Figure 1 shows the basic hierarchy of a basic HTML document.

The goal of the analysis of HTML (DOM) structure is to get a semantic structure, a structure that uses HTML tags for the purpose what they are designed to be used to, for example, <table> only for tabular data like a calendar, for divisions of a document only for generic block level element you want to style. Having a clean semantic document has a lot of benefits. It's more logical for machines to understand. That means more compatibility as old browsers and other devices like mobiles phones or PDA's will understand it. Also, it will be more logical for other machine type things to read like search engine robots giving higher search ranks. For the analysis, an open source tool (Graphviz) is used to show good or bad structure and a bad example is modified as a good one.

† 교신저자: 박정선, 경기도 용인시 처인구 명지로 116 제1공학관 516호 , 명지대학교 산업공학과
M · P: 031-330-6453, E-mail: jspark@mju.ac.kr
2011년 7월 18일 접수; 2011년 9월 24일 수정본 접수; 2011년 12월 7일 게재확정

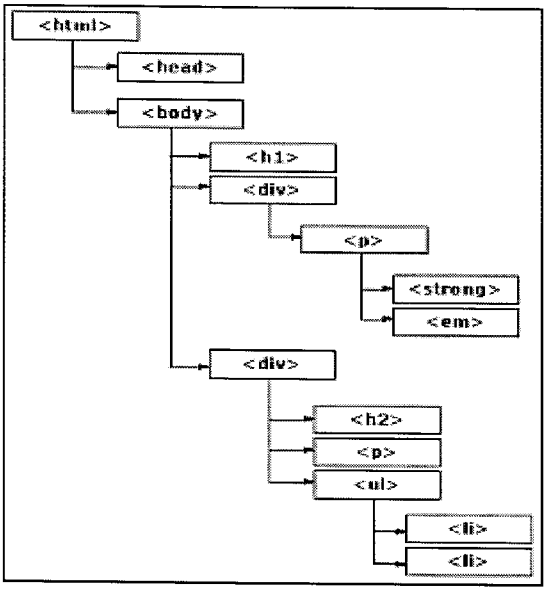


Figure 1. Hierarchy of a HTML document

2. Visualization of HTML Structure Using Tree Representation

One novel technique to analyze the HTML structure of an E commerce website is by visualizing it using tree representation where each HTML tag is represented by a node connected according to his HTML hierarchical structure.

The HTML structure can be converted in a tree graph using the following tools:

a) Perl script (html2gld.pl): This Perl script extracts the tags of a HTML documents turning in a GV file (ASCII text representation of a graph. This file describes a graph in terms of Nodes, Edges, Subgraphs, Attributes. Graphs, nodes and edges may have attributes that specify details of their appearance on the screen such as colors, sizes, shapes, etc.).

b) Graphviz (version 2.28, an open source graph visualization software): This program takes descriptions of graphs in a simple text language (GV file), and makes diagrams in useful formats, such as images and SVG for web pages, PDF or Postscript for inclusion in other documents.

Graphviz has many useful features for concrete diagrams, such as options for colors, fonts, tabular node layouts, line styles, hyperlinks.

The basic layout representations are:

- **dot** : "hierarchical" or layered drawings of directed graphs. This is the default tool to use if edges have directionality.
- **neato and fdp** : "spring model'' layouts.
- **Twopi** : radial layouts.
- **Circo** : circular layout.

A basic HTML document is represented by Figures 2 and 3:

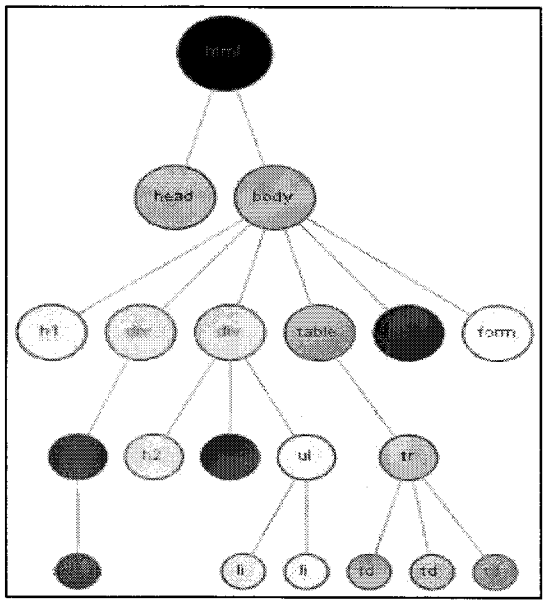


Figure 2. Visual representation of a basic HTML document in (hierarchical layout)

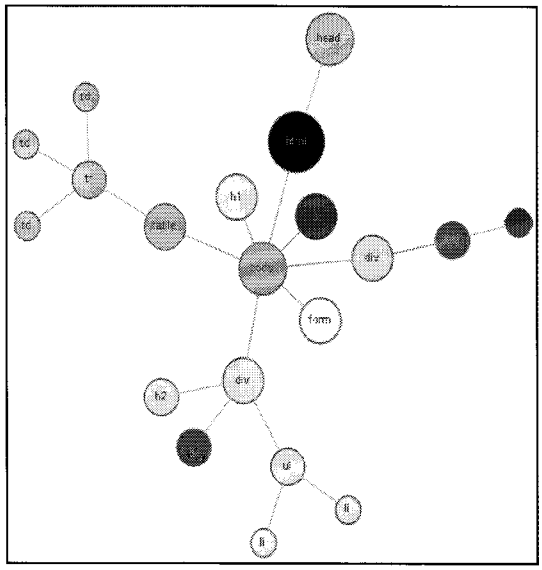


Figure 3. Visual representation of a basic HTML document in (spring layout)

A visual representation of a more complex document is shown in Figure 4.

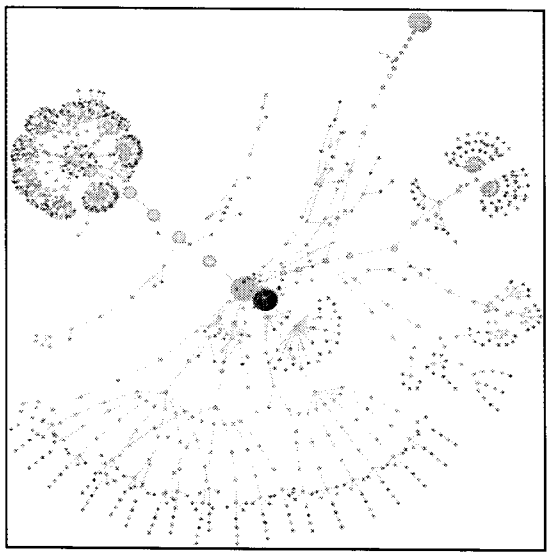


Figure 4. Visual representation of the HTML structure of Amazon.com homepage

3. Case Analysis

This analysis is based in the English website of Gmarket Inc. (english.gmarket. co.kr) accessed in May 2011. Gmarket Inc. is a famous Korean online auction and shopping mall website.

Figure 5 shows a representation of the HTML structure of the homepage of english.gmarket.co.kr using tree representation.

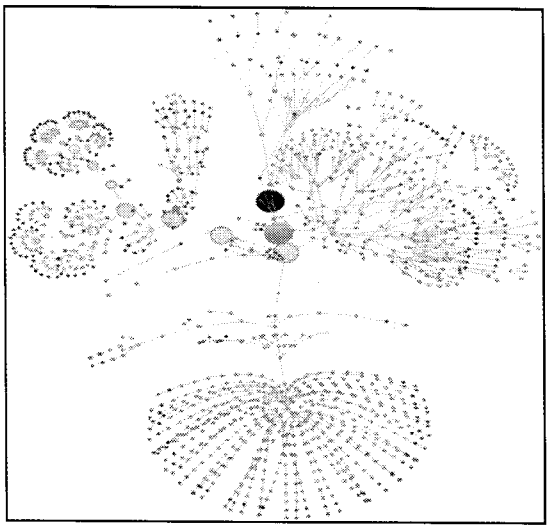


Figure 5. HTML hierarchy of english. gmarker.co.kr homepage (spring layout)

First, to analyze Figure 5 we need to know that for complex layout, table tags in general are less efficient than div tags to manage appearance. Also, tables take more HTML code and if the website is based only in tables the website is difficult to modify (the website cannot offer personalization to customer and it's not easy to maintain). However, the most important thing is to avoid the use of unnecessary tags.

The abundance of large branches with only one adjacent point usually means the tags are not being used correctly (over encapsulation, the HTML programing is not efficient and must be improved).

From the Figure 5 we can see many large branches with only one adjacent point revealing that there are table tags in excess that should be eliminate. Let's take a close look of one large branch (Figure 6).

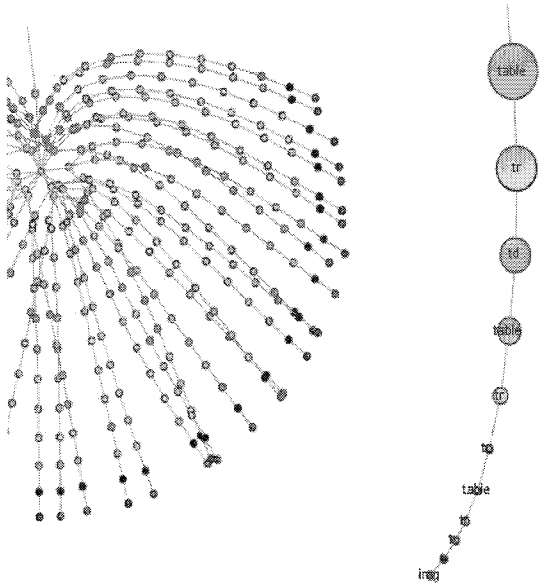


Figure 6. A close look of one large branch (spring layout)

From Figure 6 it's easy to see that there are tables in excess (there are 3 tables), what was their purpose?

To answer this question let's analyze the HTML source code of one large branch: Lines 525 537 (source code english. gmarket.co.kr homepage).

```
<table height="79" align="center"
valign="middle"> <tr><td height="79"
align="center" valign="middle">
<table width="73" border="0" cellspacing="0"
cellpadding="0">
<tr><td height="73" align="center"
valign="middle" bgcolor="#D3D3D3">
<table width="71" height="71" border="0"
cellpadding="0" cellspacing="0"
bgcolor="#FFFFFF">
   <tr><td align="center" valign="middle">

<a
href="http://english.gmarket.co.kr/challenge/neo_goods/goods.
asp?goodscode=112923056&sid=2637
6" target="_blank">

<img
src="http://goodsimg.gmarket.co.kr/goods_image2/middle_jpgi
mg/56/112923056.jpg" width="60"
height="60" border="0">

</a></td>
</tr></table></td></tr></table></td>
</tr></table>
```

The only purpose of the code was to provide a small frame for an image (a table contains the image, other table create a white space and the last table create a gray border). In this case, table tags were not necessary because only one simple div tag is necessary to create the image frame.

New code for a large branch: This is an optimized version of the code using div tags (Figure 7 shows how to look the new branch vs. its original version).

```
<div style="margin: 3px auto 3px auto;
 padding: 5px; width: 61px; border: 1px solid
 #D3D3D3; height: 61px;">

<a
href="http://english.gmarket.co.kr/challenge/neo_g
oods/goods.asp?goodscode=112923056&sid=2637
6" target="_blank">

<img
src="http://goodsimg.gmarket.co.kr/goods_image2/
middle_jpgimg/56/112923056.jpg" style="border:
 0;width: 60px;height: 60px;">

</a>
</div>
```

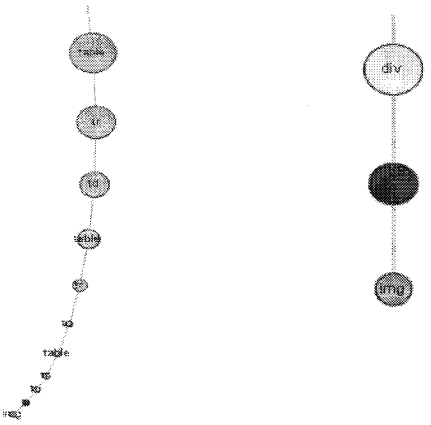
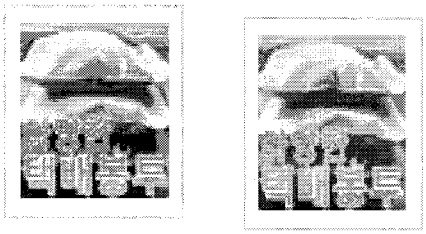


Figure 7. Visual representation of old code vs. new code of a large branch.



Output(Original Code) Output(New Code)

The above outputs of both codes are the same, but code 2 is more efficient because it required less line codes.

If we check other pages in the website we will find the same problem. Figure 8 shows the HMIL hierarchy of other web page in english.gmarkert.co.kr (Electronics > Laptops & Desktops).

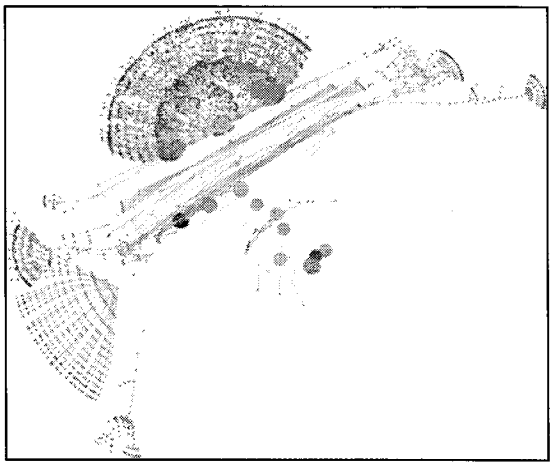


Figure 8. Visual representation of HTML hierarchy of other webpage in english .gmarker.co.kr (Electronics > Laptops & Desktops)

The analysis of Figures 5 and 8, and other web pages in the English Gmarket website revealed that this basic improvement (substituting a div tag for three table tags) can be done in all website. It will increase the speed of website and also reduce considerably the size of the HTML code.

4. Conclusions

The visualization of the HTML structure with this technique provides a good representation about how efficient and wel l-organized is the HTML programming of a web page.

The basic analysis presented in the study case and other more complex analysis can be done with this tool.

The great advantage of this tool is that developers can identify very fast potential errors in the HTML programming and improvements to have a clear semantic structure so that it will improve the speed and ranking of the E commerce website.

5. References

[1] Lee Underwood, "The HTML Hierarchy: Thinking inside the box", http://www.htmlgoodies.com/beyond/article.php/3681551/The HTML Hierarchy Thinking Inside the Box.htm

[2] Websites as graphs, 2006, http://www.aharef.info/2006 /05/websites_as_graphs.htm

[3] Oleg Burcala, "HTML as graphs: the HTML2GDL application",2009, http://www.burlaca.com/2009/02 /html2gdl graphviz/

[4] Graph Visualization Software (GraphVis), http://www.graphviz.org/

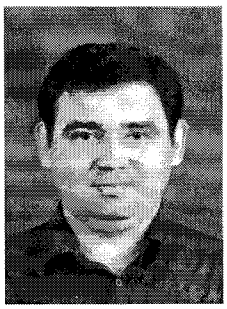
[5] Wikipedia, Gmarket, http://en.wikipedia.org /wiki /G Market

[6] http://english.gmarket.co.kr

[7] http://www.amazon.com

저 자 소 개

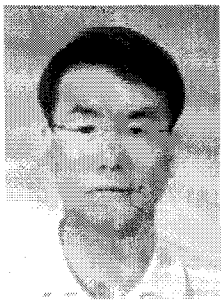
호세 밴토라



ESEN 엘살바도르에서 학사, 명지대학교에서 석사학위를 취득하였고, 관심분야는 정보의 시각화, 시각적 데이터마이닝, MIS 등이다.

주소: 경기도 용인시 처인구 남동 명지대학교 공학관 545호

박 정 선



서울대학교에서 학사, 한국과학기술원에서 석사학위를 취득하였고, 미국 텍사스주립대학교 경영학박사를 취득하였으며, 한국전산원에서 선임연구원을 거쳐 현재는 명지대학교 산업경영공학과 교수로 재직 중이다.

주소: 경기도 용인시 처인구 남동 명지대학교 공학관 516호