Transforming a Webpage to Become Standard-Compliant Through Reverse Engineering

Benfeng Chen and Vincent Y. Shen
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
May 22, 2006
Outline

• Background
  – Above 95% of Webpages are NOT Standard-Compliant

• Related Work -- HTML Tidy

• Our PURE approach
  – Algorithm description
  – Demonstration
  – Evaluation result

• Discussion
The Web Standards

- Standards-compliant webpages (http://www.webstandards.org)
  - Valid code (HTML, CSS, etc.)
  - (X)HTML for structure and content
  - CSS for presentation
Problem

>95% of Webpages are NOT Standard-Compliant

- Different rendering results in different browsers
  - Using invalid HTML code (e.g. `<A>abc<P>xyz</P>`)
- Misleading document structure
  - Using `<TABLE>` elements for layout
- Difficult maintenance and large page size
  - Using many `<FONT>` elements for text appearance
Possible Reasons

- Tolerant browsers
- Zealous authoring tools
- Copying from legacy webpages in new development
- ...

(c) 2006 Computer Science & Engineering Dept., HKUST
Benefits of Web Standards

• Small page size
  – Downloading time
  – Bandwidth

• Centralized presentation
  – Maintenance
  – Appearance customization

• Meaningful code
  – Machine-to-machine applications

• Wide accessibility
  – Different browsers

• ....

(c) 2006 Computer Science & Enigneering Dept., HKUST
Motivation

• To develop an automated tool that can transform a legacy webpage to be standard-compliant
  – CSS based presentation
  – Valid code
Outline

• Background
  – Above 95% of Webpages are NOT Standard-Compliant

• Related Work -- HTML Tidy

• Our PURE approach
  – Algorithm description
  – Demonstration
  – Evaluation result

• Discussion
Tidy

• An open source tool to fix invalid code
• Processing page’s code with rules
  – Insert missing tags
  – Adjust wrongly placed elements
  – Discard extra tags or undefined attributes/elements
  – ……

```
<a href="...">abc<p>xyz</p>
```

```
<a href="...">abc</a><p>xyz</p>
```

(c) 2006 Computer Science & Engineering Dept., HKUST
Tidy – “Table&Form” Case

```html
<body>
    Input below
    <table>
        <form action="t.cgi" name="f">
            <tr>
                <td>
                    <input size="20" name="q" value="" />
                </td>
            </tr>
        </table>
    </form>
</body>
```

Tidy

```html
<body>
    Input below
    <form action="t.cgi" name="f">
        <table>
            <tr>
                <td>
                    <input size="20" name="q" value="" />
                </td>
            </tr>
        </table>
    </form>
</body>
```

Blank line
Example of Tidy’s failure

Original

Tidy’s result

(c) 2006 Computer Science & Enigneering Dept., HKUST
Limitations of Tidy

• Tidy’s code-based approach
  – A large number of (possibly conflicting) rules
  – Ambiguous code causing rendering differences
  – No attempt to separate presentation with content
Outline

• Background
  – Above 95% of Webpages are NOT Standard-Compliant
• Related Work -- HTML Tidy
• Our PURE approach
  – Algorithm description
  – Demonstration
  – Evaluation result
• Discussion
The Reverse Engineering Approach

- PURE (webPage cleanUp through Reverse Engineering)
  - Rendering-based
  - General approach for all cases
  - What we see is what authors want
  - Generating valid code
PURE’s Approach - Overview

Legacy webpage → Webpage rendering engine → Page structure extraction

→ Layout reconstruction → Segmentation

→ Filling each box → New webpage with CSS
An Example Webpage

C:\WINDOWS\Help\Tours\htmlTour\default.htm

Welcome to Microsoft® Windows® XP Professional, the new version of Windows that brings your PC to life! Experience the best. Experience Windows XP. Windows XP really delivers! Just click a link on the right to experience Windows XP for yourself.
PURE’s Approach – Segmentation

• Extract each visible element’s position & size from a browser’s rendering engine
• Segment the page into boxes for layout reconstruction
  – Block-level elements (e.g. `<P>`, `<UL>`)

(c) 2006 Computer Science & Engineering Dept., HKUST
Example - Segmentation

Welcome to Microsoft® Windows® XP Professional, the new version of Windows that brings your PC to life! Experience the best. Experience Windows XP. Windows XP really delivers! Just click a link on the right to experience Windows XP for yourself:
PURE’s Approach - Layout Reconstruction

• Using `<DIV>` + CSS
• Two schemas
  – Absolute
  – Relative

```html
<DIV style="left:200;top:200;width:256;height:78;">
  ...
</DIV>

```
PURE’s Approach – Relative Layout

- A heuristic algorithm to construct relative layout

(a) Boxes placement

(b) First level

(c) Second level

(d) Third level

<!-- first level -->
<DIV>
  <!-- code for box A -->
</DIV>
<DIV>
  <!-- second level -->
  <!-- code for box B -->
</DIV>
<DIV>
  <!-- third level -->
  <!-- code for box C -->
</DIV>
<DIV>
  <!-- code for box D -->
</DIV>
</DIV>
Example – Relative Layout Reconstruction

Welcome to Microsoft® Windows® XP Professional, the new version of Windows that brings your PC to life! Experience the best. Experience Windows XP. Windows XP really delivers! Just click a link on the right to experience Windows XP for yourself:
Example – Layout Reconstruction
PURE’s Approach – Box Filling

• Generate code for each box
• Simple and independent
  – Easy maintenance
  – No impact on each other
Welcome to Microsoft®
Windows® XP Professional, the
new version of Windows that
brings your PC to life! Experience
the best. Experience Windows XP.
Windows XP really delivers! Just
click a link on the right to
experience Windows XP for
yourself:
Highlights of PURE Approach

- Avoid the difficulty of handling ambiguous code and guessing the author’s intention
- Avoid the difficulty of parsing the original HTML code
- Keep appearance consistent
- Use a divide-and-conquer strategy
Demonstration

- PURE prototype
Evaluation

- Top 500 websites from alexa.com
- Similarity
  - 100%: totally identical
  - 90%: very small differences
  - 80%: similar to the original, but one or two places are noticeably wrong

<table>
<thead>
<tr>
<th>Similarity</th>
<th>100%</th>
<th>90%</th>
<th>80%</th>
<th>&gt;=80% (Success)</th>
<th>&lt; 80% (Failure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>70</td>
<td>51</td>
<td>103</td>
<td>224</td>
<td>216</td>
</tr>
<tr>
<td>Rate</td>
<td>16%</td>
<td>12%</td>
<td>23%</td>
<td>51%</td>
<td>49%</td>
</tr>
</tbody>
</table>
PURE vs. Tidy

Original

Tidy’s result

PURE’s result

(c) 2006 Computer Science & Engineering Dept., HKUST
Conclusion and Discussion

• We proposed a novel reverse engineering approach to transform legacy webpages to become standards-compliant.
• We built a prototype and tested it on the top 500 websites.
• PURE's approach could be adapted to transform a webpage's layout by rearranging the boxes for mobile browsing.
Thanks

Benfeng Chen
bfchen@ust.hk
Browsing www.bokee.com in IE and Netscape (2005-12-13)