# WebAlchemist: A Structure-Aware Web Transcoding System for Mobile Devices*

Yonghyun Hwang
School of Computer Science and Engineering
Seoul National University
Seoul, Korea
semiaion@hotmail.com

Eunkyong Seo
School of Computer Science and Engineering
Seoul National University
Seoul, Korea
ilekseo@hotmail.com

Jihong Kim †
School of Computer Science and Engineering
Seoul National University
Seoul, Korea
jihong@davinci.snu.ac.kr

## ABSTRACT

We describe the design and implementation of WebAlchemist, a structure-aware Web transcoding system for mobile devices. WebAlchemist automatically converts a given Web page into a sequence of equivalent smaller Web pages with more important items of the original Web page displayed on the first transcoded page. Unlike the existing transcoding systems, WebAlchemist preserves the page layout of an original Web page as much as possible, improving the transcoding quality significantly for complex Web pages. The structure-aware transcoding heuristics of WebAlchemist extract partial but useful semantic information using an intelligent syntactic analysis. Subjective evaluation results using popular Web sites show that WebAlchemist produces satisfactory results for most Web pages tested, especially for complex Web pages.

## Keywords

Web transcoding; mobile Internet; mobile Web access; handheld devices

## 1. INTRODUCTION

With an exponential growth of mobile communications along with the pervasive use of Web in daily tasks, there is a strong need for mobile Web accesses from various handheld mobile devices. However, the current experiences of accessing Web pages from handheld devices are mostly unpleasant because of the large mismatch between the available computing resources of the handheld devices and the computing resources required for smooth viewing of the Web pages.

In solving the mismatch problem between Web contents and mobile devices, there are generally two approaches available, *manual reauthoring* [1, 2, 3, 4, 5] and *automatic reauthoring* [6, 7, 8, 9]. Since the manual reauthoring approach requires Web pages to be modified for mobile devices before they are accessed, it severely limits accessible Web pages from mobile devices. For example, during mobile searches, it is generally impossible to know the next Web page to visit, therefore, the manual reauthoring approach will not work unless user is willing to limit the search space to the reformatted Web pages only.

The automatic reauthoring approach is a better solution because it converts Web pages for a given mobile device in a fully transparent fashion to Web page developers as well as Web users. However, the existing automatic transcoding systems are quite limited; they work well with small well-structured Web pages but generate almost unusable transcoded pages for complex Web pages. The poor transcoding quality of the existing automatic transcoding systems mostly come from the fact that they ignore the semantics of a Web page.

A recent work by C. Jinlin *et al.* [10] partially considers the overall layout of a Web page during the transcoding process. However, their approach does not reflect the relative importance of various Web components; it is likely that important components are mistakenly hided behind hyperlinks. Furthermore, since their approach is based on a decision tree [10] that relies on a small set of syntactic combinations, it remains to be seen that the decision tree is general enough to support a wider range of syntactic combinations.

In this paper we describe the design and implementation of the WebAlchemist system that can generate high quality transcoded pages not only for small well-structured Web pages but also for large complex Web pages. WebAlchemist employs structure-aware transcoding heuristics as well as existing structure-unaware transcoding heuristics. We describe three structure-aware transcoding heuristics that can extract partial semantics from the bare syntactic structure of Web pages, while preserving the overall layout of Web pages as much as possible.

The rest of the paper is organized as follows. In Section 2, we give an overview of the WebAlchemist system and describe its key components including three structure-aware transcoding heuristics. We describe experimental results based on a subjective quality evaluation in Section 3. We conclude with a summary in Section 4.

*This work was supported in part by Korea Research Foundation Grant (KRF-2001-041-E00243).

†All the correspondence should be addressed to:
Prof. Jihong Kim
School of Computer Science and Engineering
Seoul National University ENG4190
Shilim-dong, Kwanak-ku, Seoul, Korea 151-742
Phone: +82-2-880-8792, Fax: +82-2-871-4912
E-mail:jihong@davinci.snu.ac.kr


*Mobile Search Workshop*, May 7, 2002, Honolulu, Hawaii, USA.
ACM 1-58113-449-5/02/0005.

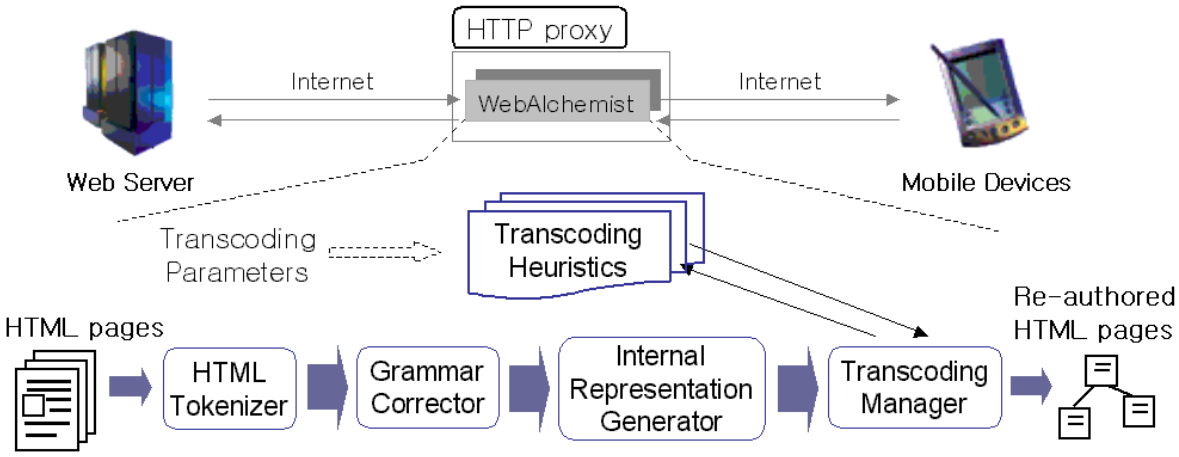



**Figure 1: An overview of the WebAlchemist system.**

## 2. THE WEBALCHEMIST SYSTEM

In this section, we give an overall architectural description of WebAlchemist and explain transcoding heuristics used in WebAlchemist.

### 2.1 Overview of the WebAlchemist System

As shown in Figure 1, WebAlchemist, being a part of a HTTP proxy server, consists of four main modules. The HTML tokenizer module classifies the content of the HTML Web page into HTML tags and non-tags. The grammar corrector module corrects HTML syntactic errors in the original HTML page, which is necessary because many Web browsers are generous to HTML syntax errors so HTML pages contain them. Since WebAlchemist transcodes based on the syntactic attribues, the grammar corrector module should fix invalid HTML syntactic errors as much as it can.

The internal representation generator module receives grammar-corrected tokenized strings as an input and converts them into a tree-based representation. The transcoding manager controls the overall transcoding procedure based on available transcoding heuristics and transcoding parameters. Once the transcoding procedure is completed, the internally represented pages are converted back to the regular HTML source format.

### 2.2 Existing Transcoding Heuristics

Before we explain structure-aware transcoding heuristics employed in the WebAlchemist system, we explain three existing transcoding heuristics used in our system, which were first introduced in [6, 11].

1. The indexed segmentation transform [6]

   This transform segments a long Web page into a sequence of small subpages. Each subpage fits into the display area of a handheld device. Small subpages are linked sequentially by hyperlinks.

2. The image reduction and elision transforms [6]

   These transforms scale down images with a predefined scaling factor, and make hyperlinks that point to the reduced images.

3. The restricted first sentence elision transform [11]

   In this transform, if a long text block is within a table structure or a text block includes a table structure, the first sentence elision transform is suppressed. The first sentence elision transform hides the whole text block behind a hyperlink except for the first sentence of the text block.

Three transcoding heuristics listed above are effective only for a small set of particular combinations of local syntactic attributes. In the next section, we propose new transcoding heuristics that additionally consider the overall layout of an original Web page.

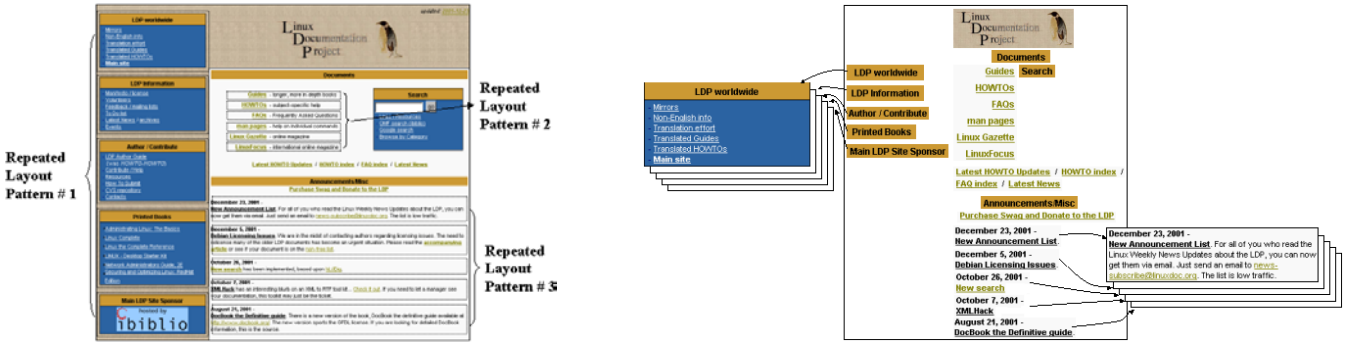
### 2.3 Structure-Aware Transcoding Heuristics

We describe three structure-aware heuristics that take advantage of common layout characteristics of complex Web pages. These heuristics are useful in transcoding complex Web pages properly.

#### *2.3.1 Generalized Outlining Transform*

Complex Web pages such as the Linux Documentation Project homepage (`http://www.linuxdoc.org`) shown in Figure 2(a) generally have a large number of repeated layout patterns. This transform identifies such repeated layout patterns in Web pages. Once repeated layout patterns are identified, they can be transcoded in a similar fashion as the original outlining transform. Figure 2(b) shows the transcoded pages by the generalized outlining transform. This transform takes advantage of the characteristics that most repeated layout structures match exactly in the fore parts while they are quite different in the trailing parts.

The repeated layout structures are detected by solving the prefix pattern matching problem on the string representation of a Web page, which is a sequentially mapped string array of HTML tags and other symbols for Web components. The mapped string array is used for constructing a string tree for the prefix pattern matching. Each edge of the string tree represents an element of the string and each node includes the number of times that the substring (that corresponds to the edges from the root to the current node) is repeated. Initially, the string tree has the root node only. When the number of elements in the string array is $N$, all the substrings whose length is between 2 and $N$-1 are in-




(a) (b)

**Figure 2: An example of the generalized outlining transform; (a) an original Web page with repeated patterns shown in dotted lines and (b) its transcoded pages using the generalized outlining transform.**

(a) (b)

**Figure 3: An example of the selective elision transform; (a) an original CNN homepage with its table outline shown and (b) its transcoded version after the selective elision transform was applied.**

serted in the string tree. The repeated substrings can be identified by a postorder traversal of the string tree.

### 2.3.2 *Selective Elision Transform*

This transform selects victim table cells and elides them while keeping table structures as much as possible. Selecting victim table cells is dependent on the syntactic attributes such as the table cell width and the font size. Whether the table cells include elided Web components or not also affects the selection of victim cells. Figures 3(a) and 3(b) show that the center table cell is wider than others and uses larger fonts. This cell is not elided under the selective elision transform, preserving the important table structure of the original CNN page as shown in Figure 3(b).

### 2.3.3 *Improved Outlining Transform*

The improved outlining transform [11] is an improved version of the outlining transform. While the original transform is applied only between the section header and following text blocks, the improved outlining transform can be used more generally when conceptually higher (more abstract) and lower (more detailed) pairs exist. In this transform, we support more syntactic combinations (e.g., the 'UL' and 'LI' tags) compared with the original outlining transform. Figures 4(a) and 4(b) illustrate the effect of the improved outlining transform for the 'UL' and 'LI' tag pairs.

## 2.4 Transcoding Manager

The WebAlchemist system supports six transcoding heuristics as stated before. The main role of the transcoding manager is to decide how these six heuristics are used for transcoding. Since transcoding heuristics require several parameters for proper operations (e.g., the display size of a handheld device and predefined threshold values for each heuristic), the transcoding manager also decides which values are used for each parameter.

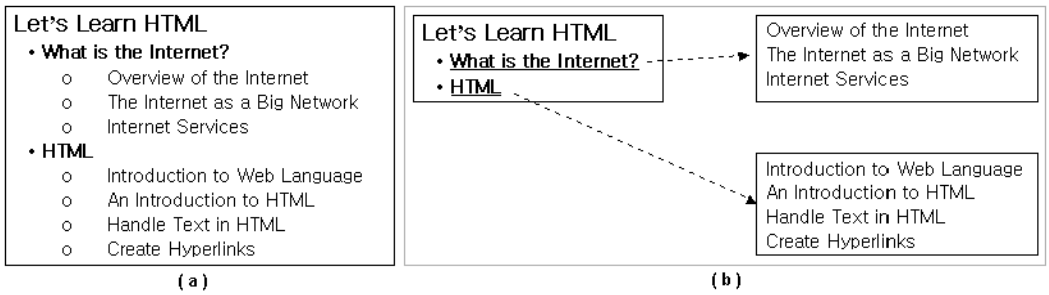



**Figure 4: An example of the improved outlining transform; (a) an original Web page with structural bullet items and (b) its transcoded pages using the improved outlining transform.**

| Name | URL (category) |
|---|---|
| chosun | `http://www.chosun.com` (newspaper) |
| cnn | `http://www.cnn.com` (broadcasting) |
| latimes | `http://www.latimes.com` (newspaper) |
| washingtonpost | `http://www.washingtonpost.com` (newspaper) |
| empas | `http://www.empas.com` (portal) |
| hotbot | `http://www.hotbot.com` |
| lycos | `http://www.lycos.com` |
| yahoo | `http://www.yahoo.com` |
| yahoo | search result of yahoo |
| mobicom | `http://www.research.ibm.com/acm_sigmobile_conf_2001` (conference) |
| nasa | `http://www.nasa.gov` (government) |
| gnu | `http://www.gnu.org` (organization) |
| lab_member | `http://davinci.snu.ac.kr/members/blublood.html` (personal homepage) |

**Table 1: Web pages used for the experiments.**

For a high-quality transcoding, each transcoding heuristic should be highly efficient. However, the order of using an individual heuristic also affects the transcoding quality significantly. Consider a Web page that consists of several repeated patterns nested in table structure. If the generalized outlining transform is followed by the selective elision transform, there are little opportunities for the generalized outlining transform to be applied.

The straightforward solution for the heuristic ordering problem is that the transcoding manager tries all the 6! combinations of heuristic orderings for a given Web page, but it is infeasible, even in the high-performance Web server, to try all the combinations of six transcoding heuristics because of the large computing power requirement and lack of effective quality metrics for automated evaluation.

Instead, we have performed extensive off-line evaluation tests using various Web pages and selected the following sequence as a default transcoding sequence for the WebAlchemist system:

**Default transcoding sequence.**
Step 1: the improved outlining transform
Step 2: the generalized outlining transform
Step 3: the selective elision transform
Step 4: the restricted first sentence elision transform
Step 5: the image reduction and elision transforms
Step 6: the indexed segmentation transform

The proposed sequence follows our intuition that preserving the overall page layout (or structure) of the Web page is an important requirement for high-quality transcoding of complex Web pages. First three transforms are effective in reducing the page size significantly while keeping the original layout of a Web page.

## 3. EXPERIMENTAL RESULTS

In order to evaluate how effective WebAlchemist is in converting complex Web pages, we have performed a subjective evaluation. 43 college students and engineers in Seoul had participated. We have used as test Web pages, 13 Web sites listed in Table 1. Transcoded pages were displayed in a simulated 320×240 display (which confirms to a typical PDA display area) for experiments. Figure 5 summarizes the results of our subjective evaluation. One of the transcoding result is shown in Figure 6. (The complete transcoding results are available in `http://davinci.snu.ac.kr/~WebAlchemist/experiments/index.html`.)

As shown in Figure 5, all the test pages received satisfactory ratings at least by 80% of the evaluation participants. Four test Web pages, lab_member, mobicom, nasa, and gnu, have received satisfactory ratings by all the participants; more than 80% of the satisfactory ratings were the *Excellent* grade. Furthermore, no test Web page was rated as *Unusable* by a single participant. One interesting observation from the subjective evaluation is that most mobile users tend to avoid horizontal scrolls as much as possible. This fact contributed the relatively poor scores of the CNN homepage. Considering the complexity of test pages (e.g., the NASA and GNU pages), the high evaluation scores demonstrate that WebAlchemist is effective in transcoding complex Web pages.

## 4. CONCLUSIONS

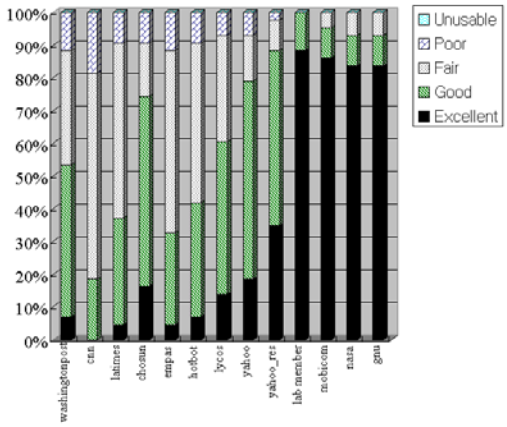



**Figure 5: Subjective evaluation results**

We have described the design and implementation of WebAlchemist, a prototype Web transcoding system, which automatically converts a given Web page into a sequence of equivalent smaller Web pages that can be displayed in a hand-held device. The unique feature of WebAlchemist is that it employs structure-aware transcoding heuristics, which can extract partial semantic information from an intelligent syntactic analysis, resulting in better transcoded Web pages.

While the current version of the WebAlchemist system produces useful HTML pages for handheld devices, it can be further improved within the automatic reauthoring framework. Our main future work is to develop more sophiscated heuristics that can extract more semantics from the syntactic analysis.

## 5. REFERENCES


[1] WAP Forum. *WAP*. `http://www.wapforum.org/`.

[2] K. Eija, A. Matti, K. Juha, M. Suvi, and L. Timo. "Two approaches to bringing internet services to WAP devices," In *Proc. 9th WWW Conf.*, 2000.

[3] F. Juliana, K. Bharat, and L. Daniel. "Webviews: accessing personalized Web content and services," In *Proc. 10th WWW Conf.*, 2001.

[4] M. Hori, G. Kondoh, K. Ono, S. Hirose, and S. Singhal. "Annotation-based web content transcoding," In *Proc. 9th WWW Conf.*, 2000.

[5] K. Nagao, Y. Shirai, and K. Squire. "Semantic annotation and transcoding," *IEEE Multimedia*, vol. 8(2), pp. 69–81, 2001.

[6] T. Bickmore, A. Girgensohn, and J. W. Sullivan. "Web page filtering and re-authoring for mobile users," *The Computer Journal*, vol. 42(6), pp. 534–546, 1999.

[7] T. Bickmore and W. Schilit. "Digestor: device-independent access to the world wide Web," *Computer Networks and ISDN Systems*, vol 29(8), pp 1075–1082, 1997.

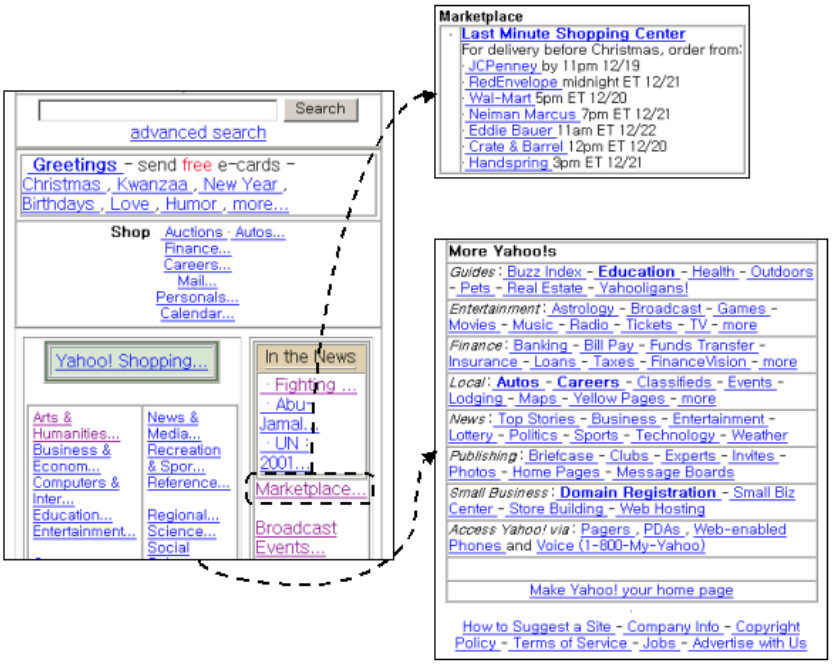
[8] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. "Seeing the whole in parts: text summarization for Web browsing on handheld devices," In *Proc. 10th WWW Conf.*, 2001.

[9] O. Buyukkokten, H. Garcia-Molina, A. Paepcke, and T. Winograd. "Power Browser: efficient Web browsing for PDAs," In *Proc. Human-Computer Interaction Conf.*, 2000.

[10] C. Jinlin, Z. Baoyao, and S. Jin. "Function-based object model towards website adaptation," In *Proc. 10th WWW Conf.*, 2001.

[11] Y. Hwang, C. Jung, J. Kim, and S. Chung. "WebAlchemist: A Web transcoding system for mobile Web access in handheld devices," In *Proc. of ITCom 2001*, 2001.


(a)




(b)

**Figure 6: A transcoding example (Yahoo homepage, `http://www.yahoo.com`); (a) an original Yahoo homepage and (b) its transcoded pages by WebAlchemist.**