ADAPTING A HYBRID-PARSING ALGORITHM INTO A CSS GENERATING TOOL

A Mini-Thesis

Presented to

The Faculty of the Computer Science Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science Major in Computer Science

By

Suñer, Carla Mae R.

Uy, Michelle Rae Q.

Viernes, Earl Denton L.

March 2004

ABSTRACT

Algorithm is a process of problem solving that will employ a mechanical sequence of steps. There are many existing algorithms at present which are designed for different purposes. On the other hand, Hyper Text Markup Language (HTML) is a text file containing small markup tags that tell the Web browser how to display a Web page. Cascading Style Sheets (CSS) specify the styles that define how to display HTML elements. It solves the problem of retyping the HTML codes for the same style to different HTML elements repeatedly. Both are tools used by the web developers in constructing the pages in the Internet. However, at present, the developers themselves are the ones responsible for creating the CSS files that are being used in the HTML documents. This study aims to build a hybrid-parsing algorithm to be adapted into a CSS generating tool. The said algorithm is developed to extract the styles from the HTML documents and to automatically place those in the CSS file. In return, the CSS file produced can be utilized by multiple HTML documents. Thus, making it easier for the web developers to maintain the uniformity of the appearance of the pages as well as separating the content from the design of the Web sites they are constructing.

Keywords:

Top Down Parsing, Infix to Postfix Conversion, Insertion Sort, Pattern Matching, CSS, HTML

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Algorithms – whether parsing, sorting, or conversion – are as much part of computers as the physical parts on which they run. Every existing application running on a computer follows a certain algorithm or a cross of different algorithms to perform their intended tasks. Naturally, in developing a new application or software that performs a specialized task, one must develop an algorithm that will carry out the task.

At the present, the growth of popularity of the Internet applications is constantly accelerating. As a result, the technologies that aid the web developers in creating these applications experience further innovations. One of the earliest technologies that serve as a tool for building these applications is the Hypertext Markup Language (HTML).

HTML contains and uses markup tags that transform into a web page and incorporate the desired styles to the web pages. However, these markup tags consist only of simple codes that do not take order and proper form much consideration. Nowadays, demands for the creation of complex web sites pressure the web designers into constructing complicated applications, giving them less time to worry about the structure of their HTML documents. With Cascading Style Sheets (CSS), developers possess the flexibility of separating

web page content from its appearance as well as use the same styles in different pages without spending much time on creating the codes.

This study involves the development of a hybrid-parsing algorithm from different existing algorithms that carries out the task of parsing an HTML document and extracting appearance-specifying tags and its adaptation into a tool that automatically generates CSS from plain HTML. As a result, the designs included in the single HTML file is transferred to a CSS document, which can be used in other HTML documents, without compromising the original appearance

1.2 Statement of the Problem

The general problem of study is how to develop a hybrid-parsing algorithm and adapt it into a CSS generating tool.

The following are the specific problems that will support and help in solving the general problem:

1. What are the existing studies and works related HTML document parsing and text extraction?
2. What is a hybrid-parsing algorithm?
3. What is a CSS?
4. What are the advantages and disadvantages of CSS?
5. What are the appearance-specifying tags and attributes in plain HTML that CSS covers?
6. How to create a grammar as a basis for identifying the HTML syntax?

7. What are the existing algorithms that the proponents use in developing the hybrid-parsing algorithm?
8. How can the proponents test the efficiency of the algorithm?

1.3 Objective of the Study

The general objective of the study is to develop a hybrid-parsing algorithm and adapt it into a CSS generating tool.

Specifically, this study seeks to:

1. Study and discuss the existing studies and works related to HTML parsing and text extraction
2. Describe a hybrid-parsing algorithm
3. Present a background on CSS
4. Identify the advantages and disadvantages of CSS
5. Identify and describe the appearance-specifying tags and attributes in plain HTML that CSS covers
6. Describe and formulate the productions of the grammar that serve as guide and basis for the identification of the HTML syntax
7. Identify, describe and discuss the existing algorithms being used in developing the hybrid-parsing algorithm
8. Develop a CSS generating tool and test the technical output

1.4 Scope and Limitation

The focus of the study is to develop a hybrid-parsing algorithm and adapt it into a CSS generating tool. The hybrid-parsing algorithm is a combination of existing algorithms and methods, modified and shaped to comply with its intended purpose. In addition, the study covers the development of a technical output, which is the CSS generating tool. This tool will produce CSS from plain HTML 4.0 documents by scanning and extracting from the HTML document the style-specifying tags. However, the study does not cover XHTML documents.

1.5 Significance of the Study

The main beneficiaries of this study are the Internet project developers that use HTML in constructing their web pages. The study will provide future thesis proponents ideas on how to merge existing algorithms and mold them into a new algorithm, which they can use for special purposes, such as text extraction or separation. The technical output will assist them in the creation of CSS from the HTML files that they have created. In addition, it will speed up their work rate by allowing them to apply the designs they have included in the single or multiple HTML files with the absence of repetitively writing the appearance-specifying tags and attributes for every web page. As CSS allows the separation of the content of the Web page from its design, the developers will find it easier to make revisions and updates on one without affecting the other.