An Imperial Investigation of Validation Problem of Top Indian Websites

Rishabh Chambial[1], Swati Sharma[2]

1 Asstt. Professor, Department of Computer Science, Sri Sai University, Palampur,India

2 Lecturer, Department of Computer Science, Sri Sai University, Palampur, India

ABSTRACT:

This paper results the report of an experimental study on the validation problem of the web pages developed by the web developers. It is found that there are very few web pages which are "VALID" and most of them are "INVALID" in any way. An invalid web page can be invalid due to many errors which are discussed in this paper. This paper uses the "W3C Validation" tool which is used to check the validations of the web pages. These web pages are of INDIA domains and are mostly used in our day to day life like facebook, orkut, rediff, irctc, hindu etc. This paper will help the standard bodies, browser developers and web page designers to work together so that the number of valid web pages continues to grow and the Web can indeed reach its full potential.

Keywords: *Error, HTML Specification, HTML Validator, W3C, Web page.*

I. INTRODUCTION

World Wide Web, is the biggest and largest information resource, which was invented in early 1990's. As it is one of the most popular sources of information, the number of website are increasing day by day. The total number of available web pages in this world is 10.58 billion on 12 September 2011 [1].

People are interested in the average size of HTML document, the average number of images in a webpage, the average size of website and so on [2] [3]. Our concern is basically about the quality of the HTML documents in today's web pages in terms of conformance to public standards.

Hypertext Markup Language (**HTML**) is the predominant markup language for web pages. HTML elements are the basic building-blocks of web pages. HTML is written in the form of HTML elements consisting of *tags*, enclosed in angle brackets (like <html>), within the web page content. HTML tags most commonly come in pairs like <h1> and </h1>, although some tags, known as *empty elements*, are unpaired, for example <img>. The first tag in a pair is the *start tag*, the second tag is the *end tag* (they are also called *opening tags* and *closing tags*). In between these tags web designers can add text, tags, comments, and other types of text-based content.

The purpose of a web browser is to read HTML documents and compose them into visible or audible web pages. The browser does not display the HTML tags, but uses the tags to interpret the content of the page.

HTML elements form the building blocks of all websites. HTML allows images and objects to be embedded and can be used to create interactive forms. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. It can embed scripts in languages such as JavaScript which affect the behavior of HTML web pages.

Web browsers can also refer to Cascading Style Sheets (CSS) to define the appearance and layout of text and other material. The W3C, maintainer of both the HTML and the CSS standards, encourages the use of CSS over explicitly presentational HTML markup. [4]

The **Markup Validation Service** is a validator by the World Wide Web Consortium (W3C) that allows Internet users to check HTML and XHTML documents for well-formed markup. Markup validation is an important step towards ensuring the technical quality of web pages;

however, is not a complete measure of Web standards conformance. [5]

The Markup Validation Service began as *The Kinder, Gentler HTML Validator*, a project by Gerald Oskoboiny. It was developed to be a more intuitive version of the first online HTML validator written by Dan Connolly and Mark Gaither, which was announced on July 13, 1994. [6]

In September 1997, Oskoboiny began working for the W3C, and on December 18, 1997, the W3C announced its *W3C HTML Validator* based upon his work. [7][8]

II. RESEARCH METHODOLOGY

This paper basically works on finding the basic percentage of errors which are found in our websites. This thing has been done by checking the links of websites on W3C validation tool which lists down the total number of errors present in ore web page and also show the category of the error in which it falls. So it is easy to find out the type of error and it will help the developers to work on these errors which are occurring in our web pages which will also help to improve the quality of the web page and also to make it a valid web page.

In this, the W3C validator [9] tool is used to validate the webpage of Indian domains which we have taken from website [10].

III. RESULTS

The results includes the nine "common" problems [11] identified by WDG. A more detailed analysis shows that some of these problems are not affecting a significant number of existing web pages.

Now further explain the most common errors below. All the results in **Figure 1** in this are based upon the results which are occurred while working on the Indian best websites on W3C validation tool.

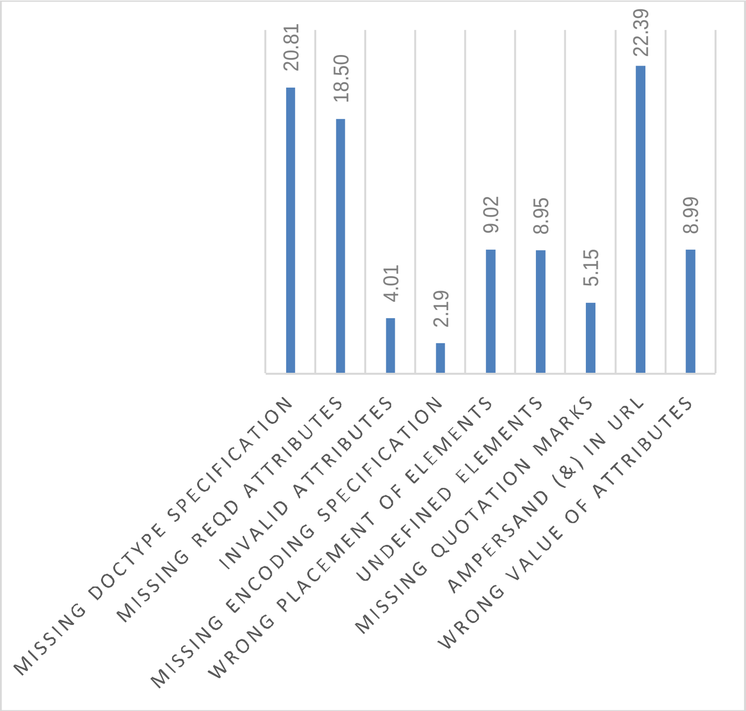

Figure 1. Most common errors

3.1 Missing DOCTYPE Specification

The DOCTYPE statement is a machine readable statement in the HTML document which specifies the structure, elements, and attributes used in that document [12]. The work shows that there are 20.81 % of the web pages omitted this important specification. Since so many web pages did not have the DOCTYPE specification, the validator automatically chose the most tolerant specification (HTML 4.01 Transitional) before proceeding to parse and analyze the HTML document.

3.2 Missing Required Attributes

The paper shows that there are about 18.50% of errors due to Missing Required Attribute. The **Figure 2** shows that while looking for the Missing Required Attributes, the web pages have about 74.88% of error which is due to "alt", 22.32% of error are due to "type", 2.33% of error are due to "action" and in last the remaining 0.47% of error is due to "content". About 18.05% of total errors are due to this "Missing Required Attributes" type.

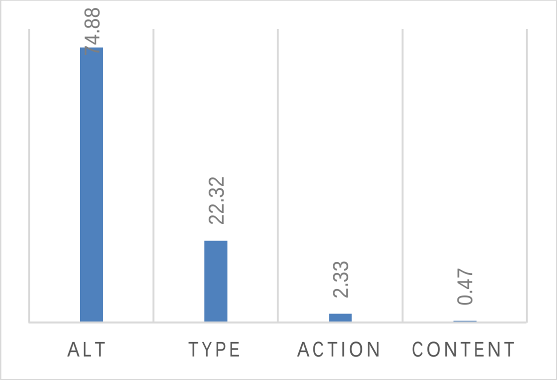

Figure 2. Missing Required Attributes errors

3.3 Invalid Attributes

The result shows that there are about 4.01% of errors due to Invalid Attributes. In addition to omitting some required attributes, web pages designers sometimes include attributes, that are not defined in the HTML standard, or defined but not for the elements they are using. **Figure 3** shows the distribution of these invalid attributes.

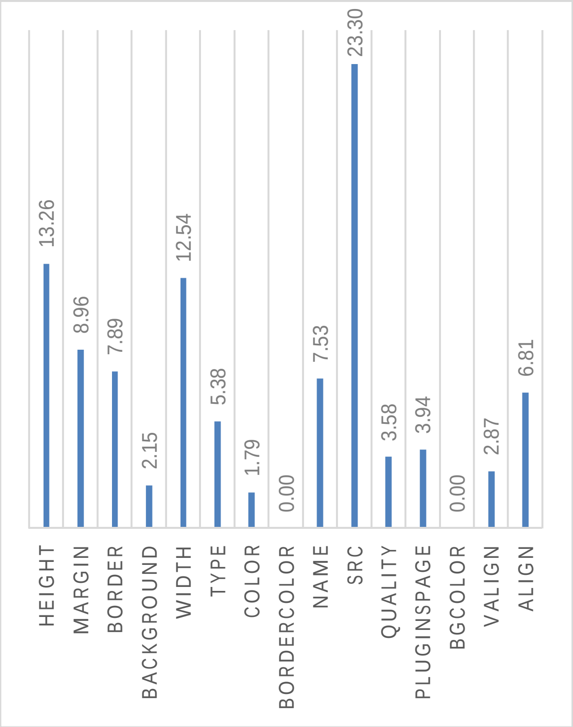

Figure 3. Invalid Attribute Errors

3.4 Missing Encoding Specification

The result shows that there are about 4.01% of errors which are occurred due to invalid attributes in our Indian websites. To improve the interoperability, each HTML document needs to specify a character set for the document, which is called the character encoding in the HTML standard. If an HTML document does not provide the encoding information, the browser may not render the character correctly on the web pages. The user must manually choose the encoding while browsing, which is inconvenient. In order to validate an HTML file without encoding specification, the validator automatically choose UTF-8 encoding that contains most known characters. However, some characters in ISO-8859- 1 (Western Europe), which is another commonly used encoding specification, are illegal or incompatible with UTF-8. If the actual encoding is ISO-8859-1 and the encoding specification is not provided in the HTML document, the validator may fail to validate the file. Therefore, in our experiment we tried to let the validator use ISO-8859-1 as the encoding to validate if the earlier attempt to use UTF-8 failed.

3.5 Wrong Placement of Elements

The results shows that there are about 9.02% of errors due to Wrong Placement of Elements. The **Figure 4** shows the distribution of wrong placement of elements errors. It is found that a maximum of 16.91% of error is of element "A" and a minimum of 0.00% of element "HR". Further the errors are being shown in the **Figure 4**

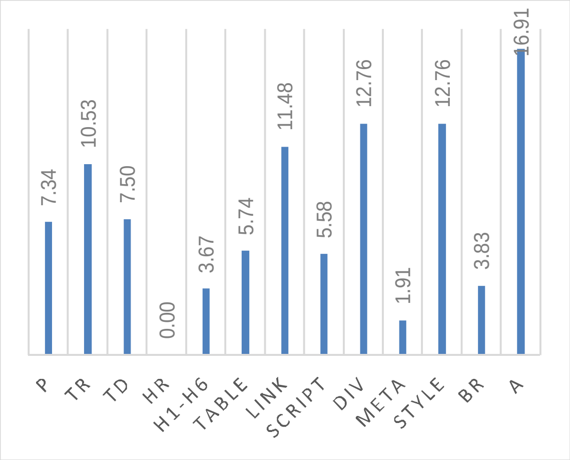

Figure 4. Wrong Placement of Elements errors

3.6 Undefined Elements

The result shows that there are about 8.95 % of errors due to undefined elements. Before the HTML standard was established some browser vendors defined elements that would be useful in creating exciting web pages. The undefined elements are those elements which are not properly defined in the HTML standards.

3.7 Missing Quotation Marks

The result shows that there are about 5.15 % of errors due to the Missing Quotation Marks. The HTML standard requires quotation marks around the literal values of attributes. It is also surprising that when Internet Explorer is used to save an HTML document using the function "Save As Web Page", it removes most of the quotation marks around literal values.

3.8 Ampersand (&) in URL

The result shows that there are about 22.39% of errors, which is maximum in the category, is due to the Ampersand (&) in URL. This type of error occurs when the URL having the ampersand (&) is used as a hyperlink in an HTML document as the "&" is a reserved symbol in HTML that is the beginning of the entity. If it has to used as a hyperlink then it should be changed to "&amps" in the hyperlink.

3.9 Wrong Value of Attributes

The result shows that there are about 8.99% of errors which are due to wrong value of attributes. In the HTML standard some attributes have a specific range of values. **Figure 5** shows the distribution of such errors.

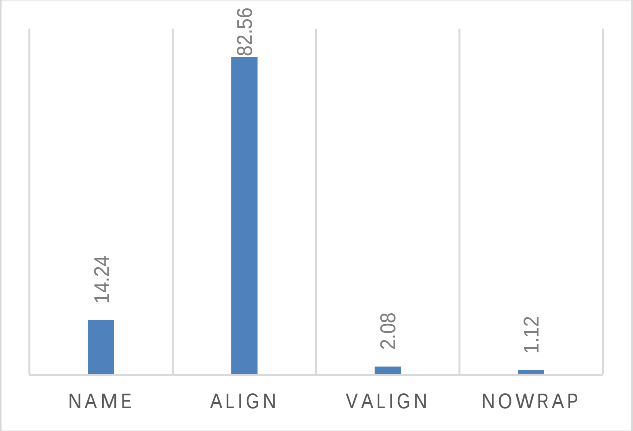

Figure 5. Wrong Value of Attributes Error

This part shows that the maximum error is done by the attribute "align" which is about 82.56% and falls highest in this category. The other attributes like "name", "valign" and "nowrap" also comes under this category.

IV. CONCLUSION

The validation problem in existing webpages have drawn more and more attention because of the increasing trend of Web communication moving from a human-to-computer process to a computer-to-computer process. In this paper, it is concluded that there is major problem in the validation problem of the current web pages. It works on the *9 major errors* which are found in almost any web page. The experiment is done upon *50 Indian websites.* It is disappointing to find that none of the webpage in the list is valid according to this test results. It is believed that the findings of this paper are useful to the standard bodies and authoring tool vendors, as well as to the Web application developers and web page designers. As due to diversity in validation problem, it is very difficult to fix all the errors in existing web pages. It is believe that due to the deliberate efforts and with the collaboration of standard bodies, web page designers and vendors, the validation problems can be reduced till some extent and the Web can reach its full potential.

REFERENCES

[1] http://www.worldwidewebsize.com/

[2] Lawrence, Steve, and Giles, C. Lee, Accessibility of information on the web. Nature 400 (July 1999), 107-109

[3] Lee, D.C., and Midkiff, S.F. A sample statistical characterization of the world wide web. Proceeding of Southeastcon '97. 'Engineering New Century' (April 1997), 174-178.

[4] http://en.wikipedia.org/wiki/HTML

[5] "About the W3C Markup Validation Service". *W3C Markup Validation Service*. World Wide Web Consortium. http://validator.w3.org/about.html. Retrieved 2011-05-20.

[6] Connolly, Dan (1994-07-13). "ANNOUNCE: HaL Interactive HTML Validation Service". *www-html mailing list*. http://lists.w3.org/Archives/Public/www-html/1994Jul/0015. Retrieved 2008-06-28.

[7] Oskoboiny, Gerald (2003-03-22). "History of the Kinder, Gentler HTML Validator". http://impressive.net/people/gerald/1996/ugweb/validate/changes.html. Retrieved 2008-06-28.

[8] http://en.wikipedia.org/wiki/W3C_Markup_Validation_Service

[9] HTML Validation Tool, http://validator.w3.org/

[10] http://www.bestindiansites.com/ 29 August, 2011

[11] Web Design Group, Common HTML Validation Problem.

http://htmlhelp.com/tools/validator/problems.html

[12] W3C, Don't forget to add a doctype.

http://www.w3c.org/QA/Tips/Doctype.