Volume 9, No. 1, January-February 2018



International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DESIGN AND DEVELOPMENT OF SIMULATION TOOL FOR TESTING SEO COMPLIANCE OF A WEB PAGE – A CASE STUDY

Dr. P.G. Naik Professor Dept. of Computer Studies CSIBER, Kolhapur, India Dr. R.S. Kamath Associate Professor Dept. of Computer Studies CSIBER, Kolhapur, India Dr. S.S. Jamsandekar Assistant Professor Dept. of Computer Studies CSIBER, Kolhapur, India

Abstract: Efficient search engines enable the end user to avail the target information as early as and as accurately as possible. Search engine optimization conforms to the act of optimizing the webpage in particular and a website in general so as to acquire a highest ranking in search engine results. The ranking is based on some pre-defined criteria which SEO encapsulates for generating right signal to the search engine. The basic methodology adopted for this purpose primarily focuses on enriching the website content and improving readability by making the web pages more search engine friendly. Several optimization tools exist for rendering the website search engine friendly. But all these tools operate like a black box where the internal details are not immediately revealed to an end user. To facilitate this, in the current work the authors have designed and developed a simulation tool for testing w3C compliance which is a necessary condition for making a webpage search engine friendly. The browsers do not complain against the violation of W3C compliance rules, however such violations do result in rendering the web page invisible to the search engine which in turn results less network traffic flow to the website. Among other things W3C and CSS compliance are the most basic rules which a website designer cannot afford to ignore. In the current work, the tool is designed and developed for the purpose which accepts an HTML document and computes its W3C compliance score. The tool meticulously analyzes the HTML code and generates several reports conforming to violation of rules, if any. Further, the tool has the provision for converting the document into W3C compliance document based on the specified ruleset. Finally, the tool has been tested by applying it to the case study of a hypothetical organization.

Keywords: CSS, Digital Marketing, HTML, JavaScript, Off-site Optimization, On-site Optimization, W3C Compliance

1. INTRODUCTION

In the era of Internet and worldwide web, digital marketing has emerged as the most promising and powerful trend in marketing which focuses on retaining the existing customers and acquiring the new customers based on the digital content. Academic institutions are not an exception to this. Due to the extensive competition among educational institutions for acquiring quality students, institutions are also adopting these advanced cutting edge technologies in establishing their brand names and reaching the aspirants across the globe. There is vast amount of digital content already present on Internet and more content is constantly generated at an exponential rate. Due to an intense competition, every organization intends to see itself at the top of the content generated by the search engines. It has been predicted that in the coming decade, in this entire competition only those organizations which quickly adopt themselves to these cutting edge technologies will survive which gives them an upper hand over their competitors and will consequently be the emerging winners conforming to the Darwinian principle of survival of the fittest. Currently most of the academic institutions are adopting Search Engine Optimization (SEO) for rendering their sites search engine friendly thereby sailing up to higher positions in search engine results.

Search engine optimization is a digital marketing methodology dealing with the measures taken towards improvement of a ranking of a website in search engine results thereby resulting in an increase in page views. SEO optimization can be achieved with the help of a set of rules pertaining to keyword selection, image optimization, focusing on rich web content to name a few which in turn renders the website search engine friendly. There exist a fistful of freelancing companies providing these facilities. SEO focuses on improving web visibility through a set of specific keywords. SEO in essence is a guiding theory for increasing web traffic. Keyword is the most crucial factor which plays a key role in influencing search engine results to a large extent. The generic optimization principle namely, early termination technique is employed by majority of search engines which involves computation of results based on indexing and employing pruning techniques which reduces the huge computational task involved in exhaustive traversal.

The most challenging task for the search engine is the tremendous workload to be handled by it and on an average the typical search engine needs to address hundreds of millions of queries on per-day basis involving billions of documents. Search engine addresses this issue by involving millions of machines distributed over multiple data centers with an intention to reduce query processing cost. The SEO techniques primarily fall into two distinct categories

- On-site optimization
- Off-site optimization

On-site optimization demands meticulous attention during the phase of website development targeting domain names, file names, directory structure, title, keyword selection and keyword density. Another area which requires meticulous attention is keyword consistency within the headings, anchor tags, alternate tags, descriptions etc.

Off-site optimization on the other hand focuses on inclusion of a website on various social networking sites for increasing network traffic to the website, availing a reputed web hosting service, blog publishing, link building etc. The parameter employed for measuring the off-site optimization is Click-Through-Rate (CTR) of a website which is computed by dividing the number of times the page containing the link is displayed by number of times the link is clicked. For example if the page containing the link is clicked 10,000 times and the link is clicked 100 times, then CTR for the page is (100/10000)*100 % = 1%.

SEO optimization is based on four optimization techniques: structure optimization, keyword optimization, content optimization and link optimization. In the current work, authors focus on content optimization based on few key rules and validate the HTML document against those rules. Based on the validation process a compliance score is assigned to the document. The tool also contains the provision for converting invalid HTML documents into their valid counterparts. Currently, the tool operates only with a few set of rules which can easily be extended to incorporate additional rules.

2. LITERATURE REVIEW

With the rapid evolution of information technology and explosion in the number of websites, SEO technology is continuously gaining a tremendous importance and is catching the attention of every website developer. SEO techniques focus on few parameters pertaining to keyword selection, back linking, rich dynamic content and target towards gaining a better ranking in the search engine results [1, 2]. Website promotion is just a single application of SEO technology. The application of search engine optimization is many fold. Apart from improving the page ranking, there are numerous applications which exploit search engine optimization techniques such as spamming application, six sigma management application for which SEO provides a conceptual background [2 -4]. The guidelines for building search engine friendly websites resulting in optimization of search queries have been presented [5-7]. Further, the authors present a review of search engine optimization tool as a part of digital marketing regime. Different categories of SEO tools are highlighted to aid website promotion. Search engine filtering system based on two-tier link extractor is presented to decrease the traffic of irrelevant pages in search result [8]. Search engine optimization based on soft computing technique is presented by Wang et.al [9] where a back propagation learning method of a neural network is employed for speedy retrieval of data from the web. Web caching based on semantic web technique is employed for designing a cluster of search engines with a goal to reduce load on the web server and latency time [10]. The different algorithms, lowest relative value algorithm, LUV algorithm and least weighted usage algorithm are employed to study their relative merits and demerits.

Research Gaps

Most of the work on search engine optimization techniques focus on on-site and off-site optimization techniques and primary focus is on keyword optimization. There are no simulation tools available for revealing the basic operation of these optimization tools. To account for this the current research focuses on design and development of simulation tool for testing SEO compliance of a website. The scope of the tool is rather too restricted to account for only the partial content optimization. The paper provides the first hand information for new researchers in this area.

3. CONCEPTUAL MODEL DESIGN

A. Problem Definition

XYZ institute is a leading management institute in south western Maharashtra accredited by NACC with "A" grade. The institute has in-house website development team which is actively involved in maintenance of the website. The management of the institute is interested in improving the website ranking in search engine queries.

B. Proposed Solution

The problem can be addressed at two levels.

As a first alternative, hiring a freelancer, who works for 25% cheaper than SEO company. But at the same time it can be risky and time consuming as freelancer many not be aware of the specificwebsite strategy and may struggle to understand the exact needs. Finding a good freelancer is time consuming and is more often a trial and error process.

Current Status

There are plethora of free and open source tools available on Internet which enable the analysis of the website content and give SEO ranking of the website. The tools generate the report pertaining to

- W3C Compliance Test
- CSS Compliance Test
- Keyword Analysis
- Broken links analysis

and suggest improvements for making the website SEO affine.

As a second alternative, the in-house tool can be designed and developed for verifying W3C and CSS Compliance which can then be input to available tools for improving it further. This enables the website to rank top in search engine results thereby improving the volume of traffic towards website by search engine. This demands the technical expertise which may be expected from XYZ institution due to the availability of in-house website development team.

In a simpler manner, at a first level search engine optimization can be achieved by targeting search terms which commonly include title, description, keywords, headings and alternate text. Few prominent rules for SEO are listed below:

C. Few Prominent Rules for SEO

Optimization rules exist on HTML filename and several HTML elements such as title, anchor element etc.

- Matching filenames with a page title
- Using hyphens rather than keyword for separating keywords.
- File size restriction to 101K, since text above it is chopped by most of the search engines.
- No use of frames and JavaScript embedded in HTML
- Usage of more text than HTML elements
- Using meta tag with the attribute "name" set to "keywords" and the "content" attribute listing about 25-30 keywords.

- Using Meta tag with the attribute "name" set to "description" and the "content" attribute listing about 150 characters.
- Limiting title size to maximum 9 words or 60 characters.
- Adding title to anchor element.

D. W3C Compliance Test

Browsers do not consider any violation against W3C compliance but search engines do. On useful link to validate the webpage is

validator.w3.org

Rules of W3C Compliance

- 1. Every tag must be closed explicitly. Empty tag must use terminating slash.
- 2. The head tag is mandatory.
- 3. All tags must be in lower case.
- 4. All attributes must be enclosed in double quotes.
- 5. Form tags cannot be nested.
- 6. Use entity references wherever necessary.
- 7. CSS rules should be in lower case.

E. Application Architecture

The application architecture and operation of SEO analysis tool is presented in Figure 1.



Figure 1. Application Architecture of SEO Analysis Tool

The various steps involved in testing SEO compliance are depicted below:

- 1. Accept a webpage from the end user.
- 2. Check no. of significant rules (A rule is considered to be significant if the corresponding element exists in an input web page)
- 3. Distribute the percentage among the significant rules equally.
- 4. Analyze the webpage and generate SEO compliance report.
- 5. Modify the webpage to render it more SEO friendly.

The tool operates by accepting an HTML document from an end user. HTML document is then analyzed to check the presence of some prominent tags. Seven rules for W3C compliance stated above are considered. Out of the seven rules stated above, some rules may be insignificant under the context based on the structure of HTML document input by an end user. The

primary focus is in locating un-terminated tags, unquoted attributes, missing head element etc. The score is equally distributed among the tags present and finally the SEO compliance score for the current ruleset is computed.

F. Implementation of Rules

All the seven rules carry equal priority and the percentage score reserved for each rule is given by,

Percentage = 100 / N

where, N is the number of significance rules.

The net score is computed using the formula

 $\begin{array}{c} \sum S_i \\ \text{where } 1 <= i \ <= N \end{array}$

The computation of a score for first four rules is depicted below:

Implementation of Rule 1:

The rule for the matching tags (un-terminated tags) is,

S = E + T

where, S, E and T refer to the no of start tags, end tags and terminating tags, respectively.

The score awarded after testing rule 1 is Score₁ = (Percentage) * (E+T) / S

Implementation of Rule 2:

Implementation of rule 2 is trivial since the full/zero score is awarded based on the presence/absence of a head node.

Implementation of Rule 3:

Score₃= (Percentage) * L/T where T has the same meaning as above and L is the number of tags in lower case.

Implementation of Rule 4:

Score₄= (Percentage) * Q / A

where Q is the total number of quoted attributes and A is the total number of attributes.

The final SEO compliance score is then given by,

 $S=S_1+S_2+\ldots\ldots+S_7$

4. RESULTS AND DISCUSSIONS

The model proposed above is implemented in Visual Basic 6 and is applied to the case study of XYZ educational institute stated above. The partial source code for implementation of tool is listed in Appendix A. The tool is tested for different test cases to check its validity and scope. The tool is also compared with the markup validation service provided by W3C by employing the website validator.w3c.org. The results are depicted in Figure 2(a) - Figure 2(d)

→ O validator.w3.org/check				🖾 🕁	=	Ø	۵
Markup Valida Check the markup (HTML.	ation Service						
	Jump	To: Notes and Potential	ssues Validation Output				
	Errors found w	hile checking this docu	ment as HTML 4.01 Transitional!				
Result:	3 Errors, 4 warning(s)						
File :	Use the file selection box above if you	Browse wish to re-validate the uploaded file	seo.html				
Encoding :	utf-8		(detect automatically)				
Doctype :	HTML 4.01 Transitional		(detect automatically)				
Root Element:	html						
mozilla	The W3C validators	are developed with assistanc Donate and help	e from the Mozilla Foundation, and supported by communi us build better tools for a better web.	ty donations.	0	Ci Flattr	
Options							
Show Source	Show Outline	List Messages Sector	quentially O Group Error Messages by Type				
□ Validate error pages	Verbose Output	Clean up Markup	with HTML-Tidy				
Holp on the options is available.				Activate Win	Revalid	late	

Dr. P.G. Naik et al, International Journal of Advanced Research in Computer Science, 9 (1), Jan-Feb 2018,323-340

alidation Output; 3 Errors Line 1, Column 1: no document type declaration: implying " HTML SYSTEM " Line 1, Column 1: no document type declaration: implying " HTML SYSTEM " Line 3, Column 1: no document type ("DOCTYPE") declaration. The Validater has blod to validate with a failtack DTD, but this is quite likely to be incorrect messages. It is highly recommended that you insert the proper DOCTYPE declaration in your document – instructors for doing this are given above – and it is necessary to t to be valid. Line 2, Column 7: document type does not allow element "BODY" here Line 2, Column 7: document type does not allow element "BODY" here Line 2, Column 7: document is the use of XiTML strike it is not allowed. This could mean that you have incorrectly nested elements – such as a "style" element in the "bod" that everlap (which is not allowed). One common cause for this error is the use of XiTML syntax in HTML documents. Due to HTML's rules of implicitly dosed elements, this error can create cascading effect meter and "trik" in the "bod" section of a HTML document may cause the parser to infer the end of the "head? section and the beginning of the "body" section (where "link" and	t and will generate a large number of incorrect error have this declaration before the page can be declared by section instead of inside "head" - or two elements s. For instance, using XHTML's "self-closing" tags for
	t and will generate a large number of incorrect error have this declaration before the page can be declared by section instead of inside "head" – or two elements s. For insteame, using XITMU's "self-closing" tags for
Une 1, Column 1: no document type declaration; implying " <idoctype html="" system="">" UNKE: UNKE: Une 1, Column 1: no document type ('DOCTYPE') declaration. The Validater has tried to validate with a failback DTD, but this is quite likely to be income messages. It is highly recommenced that you insert the proper DOCTYPE' declaration in your document — instructors for doing this are given above — and it is necessary to h to be valid. Une 2, Column 7: document type does not allow element "BODY" here Sums > The element named above was found in a context where it is not allowed. This could mean that you have incorrectly nested elements — such as a "style" element in the "bod' hat coverage (which is not allowed). De common cause for this error is the use of XrifML syntax in HTML documents. Due to HTML's rules of implicitly closed elements, this error can create cascading effect "meta" and "ins" in the "head" section of a HTML, document may cause the parser to infer the errol of the relation of the "body" section (where "link" and "ins" in the "head" section of a HTML document may cause the parser to infer the errol of the "read" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "ins" in the "head" section and the beginning of the "body" section (where "link" and "link" in the "head" section and the beginning of the "body" section (where "link" and "link" in the "</idoctype>	t and will generate a large number of incorrect error have this declaration before the page can be declared by section instead of inside "head" - or two elements s. For instance, using XrTML's "self-closing" tags for
INTRUS The checked page did not contain a document type ('DOCTY'PE') declaration. The Validator has tried to validate with a fallback DTD, but this is quite likely to be income massages. It is highly recommenced that you inset the proper DOCTY'PE' declaration in your document — instructors for doing this are given above — and it is necessary to to be valid. Line 2, Column 7: document type does not allow element "BODY" here REON > The denement mored above was found in a context where it is not allowed. This could mean that you have incorrectly nested elements — such as a "style" element in the "bod that contage (which is not allowed). Ore common cause for this error is the use of XHTML syntax in HTML documents. Due to HTML's rules of implicitly dosed elements, this error can create cascading effect media' and "tink" in the "head" section of a HTML, document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" a	et and will generate a large number of incorrect error have this declaration before the page can be declared by "section instead of inside "head" - or two elements s. For instead of inside "head" - or two elements
The checked page did nd contain a document type (DOCITYPE') declaration. The Validator has tried to validade with a failback DTD, but this is quite likely to be incorrect messages. It is highly recommended that you inset the proper DOCITYPE's declaration in your document — instructors for doing this are given above — and it is necessary to to to be valid. J Line 2, Column 7: document type does not allow element "BODY" here BODY > The element normed above was found in a context where it is not allowed. This could mean that you have incorrectly nested elements — such as a "style" element in the "bod" that everlap (which is not allowed). One common cause for this error can create cascading effect "meter" and "tirk" in the "head" section of a HTML, document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" and "tirk" and "tirk" in the "head" section of a HTML, document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" and "tirk" and "tirk" in the "head" section of a HTML document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" and "tirk" and "tirk" in the "head" section of a HTML document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" and "tirk" and "tirk" in the "head" section and the beginning of the "body" section (where "link" and "tirk" and tirk" and tirk" and tirk a toric and tirk and tirk" and tirk	t and will generate a large number of incorrect error nave this doclaration before the page can be declared have this doclaration before the page can be declared by section instead of inside "head" - or two elements s. For instance, using X0TINL's "self-closing" tags for an instance.
Line 2, Column 7: document type does not allow element "BODY" here Line 2, Column 7: document type does not allow element "BODY" here Line 2, Column 7: document is a context where it is not allowed. This could mean that you have incorrectly nested elements – such as a "style" element in the "bod that everlag (which is not allowed). Ore common cause for this error is the use of XITML syntax in HTML documents. Due to IITML's rules of inclicitly dosed elements, this error can create cascading effect 'meta' and "ins' in the "head" section of a HTML document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" and	h/" section instead of inside "head" or two elements s. For instance, using X0TML's "self-closing" tags for
CBOD® > The dement nermed above was found in a context where it is not allowed. This could mean that you have incorrectly nested elements such as a "style" element in the "bod that overlap (which is not allowed). One common cause for this error is the use of XHTML syntax in HTML documents. Due to HTML's rules of implicitly closed elements, this error can create cascading effect "meta" and "trik" in the "head" section and the beginning of the "body" section (where "link" and "trik" and "trik" in the "head" section and the beginning of the "body" section (where "link" and "trik" and "trik" in the "head" section and the beginning of the "body" section (where "link" and "trik" and "trik" in the "head" section and the beginning of the "body" section (where "link" and "trik"	ly" section instead of inside "head" - or two elements s. For instance, using XHTMU's "self-closing" tags for
that overlage (which is not allowed). One common cause for this error is the use of XHTML syntax in HTML documents. Due to HTML's rules of implicitly closed elements, this error can create cascading effect "meter" and "trix" in the "head" section of a HTML document may cause the parser to infer the end of the "head" section and the beginning of the "body" section (where "link" an	s. For instance, using XHTML's "self-closing" tags for
	val metal are not allowed; nence the reported error).
Line 7, Column 7: end tag for "HTML" which is not finished	
/IITNE > Most likely, you mested tags and closed them in the wrong order. For example shot acceptable, as must be closed before Acceptable mesting is 	p> <arr></arr>
Another possibility is that you used an element which requires a child element that you did not include. Hence the parent element is "not finished", not complete. For instance child element, lists require appropriate ist items (and < and and 	, in HTML the <head> element must contain a <title> Activate Windows</title></head>
	Go to Settings to activate Windo
validator.w3.org/check	W X = M
This document was Tentatively checked as HTML 4.01 Transitional	
Security Test Parts and American A	
Result: Tentatively passed, 4 warning(s)	
Result Tentatively passed, 4 warning(s) Browse	
File: Browse Use the file selection box above if you wish to re-validate the unloaded file seo2 html	
Result: File: Browse Use the file selection box above if you wish to re-validate the uploaded file seo2 html Use the file selection box above if you wish to re-validate the uploaded file seo2 html Encoding: utf-8 (detect automatically)	
Result: Ternatively passed, 4 warring(s) File: Browse Use the file selection box above if you wish to re-validate the uploaded file seo2.html Encoding: utf-8 (detect automatically) Doctype: HTML 4.01 Transitional (detect automatically)	
rentatively passed, 4 warring(s) File:	
Result: rentatively passed, 4 warring(s) File:	
Result: Fentatively passed, 4 warring(s) File:	ommunity donations.
Result Tentatively passed, 4 warning(s) Browse Browse Use the file selection box above if you wish to re-validate the uploaded file seo2 html Encoding: utf-8 (detect automatically) Doctype: HTML 4.01 Transitional (detect automatically) Root Element: html The W3C validators are hosted on server technology donated by HP, and supported by or Donate and help us build better tools for a better web.	ommunity donations.
Result: Tentatively passed, 4 warning(s) File: Browse Use the file selection box above if you wish to re-validate the uploaded file seo2 html Encoding: Ut-8 (detect automatically) Doctype: HTML 4.01 Transitional (detect automatically) Root Element: html The W3C validators are hosted on server technology donated by HP, and supported by or Donate and help us build better tools for a better web.	community donations.
Result in Tentatively passed, 4 warning(s) File: Browse Use the file selection box above if you wish to re-validate the uploaded file seo2.html Colspan="2">Colspan="2" Colspan="2">Colspan="2" HTML 4.01 Transitional Colspan="2">Colspan="2" Root Element: html The W3C validators are hosted on server technology donated by HP, and supported by on Donate and help us build better tools for a better web. Donate Conate	community donations. ර ල් flute
Result: Tentatively passed, 4 warning(s) File: Browse Use the file selection box above if you wish to re-validate the uploaded file seo2.html Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2">Colspan="2"Colspan="2	community donations.
Result Tentalively passed, 4 warning(s) Browse Browse Use the file selection box above if you wish to re-validate the uploaded file seo2 html Colspan="2">Colspan="2" Colspan="2">Colspan="2" HTML 4.01 Transitional Colspan="2">Colspan="2" Colspan="2">Colspan="2" Colspan="2">Colspan="2" HTML 4.01 Transitional Colspan="2" Colspan="2" <thcolspan="2"< th=""> Colspan="2" <thc< td=""><td>community donations. ្រី កីរវង</td></thc<></thcolspan="2"<>	community donations. ្រី កីរវង
Result Perifatively passed, 4 warning(s) Browse Browse Use the file selection box above if you wish to re-validate the uploaded file sec2.html Encoding: utf-8 (detect automatically) Doctype: HTML 4.01 Transitional (detect automatically) Root Element: html The W3C validators are hosted on server technology donated by HP, and supported by or Donate and help us build better tools for a better web. Dotter web. Dptions Show Source Show Outline List Messages Sequentially Group Error Messages by Type Validate error pages Uverbose Output Clean up Markup with HTML-Tidy	community donations. C Flats

Validation Outpu	ut: 4 Warnings	
Below is a list of the	a warning message(s) produced when checking your document.	
Å Line 8, Colum	nn 7: NET-enabling start-tag requires SHORTTAG YES	
<hr/>		
For the current doo interpretation is trig compatible *self-clo	ccurrent, the validator interprets strings like <eco></eco> according to legacy rules that break the expectations of most authors and thus cause confusing warnings and error messages from the validator. This iggored by HTML 4 documents or other SGML-based HTML documents. To avoid the messages, simply remove the "/" character in such contexts. NE: If you expect <pco></pco> to be interpreted as an XML- losing" tag, then you need to use XHTML or HTML5.	
This warning and re value.	related errors may also be caused by an unquoted attribute value containing one or more "". Example: #3Cc/a>. In such cases, the solution is to put quotation marks around the	
Congratulations	1	
Congratulations The uploaded docum of it. In other words, these changes. The	s ment "seo2.html" was <i>tentatively</i> checked as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal , the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you me parser implementations we used for this check are based on <u>OpenSP</u> (SGML/XML).	validatic ake
Congratulations The uploaded docum of it. In other words, these changes. The Validating CSS sty if you use CSS in you	s ment "seo2 html" was <i>tentatively</i> checked as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal , the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you m e parser implementations we used for this check are based on <u>OpenSP</u> (SGML/XML). <i>Yel Sheets</i> pur document, you can check it using the W3C CSS Validation Service.	validatio ake
Congratulations The uploaded docum of it. In other words, these changes. The Validating CSS Sty f you use CSS in yo	ment "seo2.html" was <i>tentatively</i> checked as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you m the parser implementations we used for this check are based on <u>OpenSP</u> (SGML/XML).	validatio ake
Congratulations The uploaded docum of it. In other words, these changes. The Validating CSS sty f you use CSS in you	s ment "seo2.html" was <i>tentatively</i> checked as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal , the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you m ie parser implementations we used for this check are based on <u>OpenSP</u> (SGML/XML). ty/e Sheets our document, you can check it using the W3C <u>CSS Validation Service</u> .	validatic ake
Congratulations The uploaded docum of it. In other words, these changes. The Validating CSS Sty f you use <u>CSS</u> in yo	The document would validate as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you m is parser implementations we used for this check are based on <u>OpenSP</u> (SGML/XML). tyle Sheets our document, you can check it using the W3C <u>CSS Validation Service</u> . Home About News Docs Help & FAQ Feedback Contribute	validatio ake
Congratulations The uploaded docum of it. In other words, these changes. The Validating CSS sty if you use CSS in yo	s ment "seo2 html" was <i>tentatively</i> checked as HTML 4.01 Transitional. This means that with the use of some fallback or override mechanism, we successfully performed a formal the document would validate as HTML 4.01 Transitional if you changed the markup to match the changes we have performed automatically, but it will not be valid until you m te parser implementations we used for this check are based on OpenSP (SGML/XML). tyle Sheets our document, you can check it using the W3C CSS Validation Service. Home About News Docs Help & FAQ Feedback Contribute This service runs the W3C Markup Validator, y <u>1.3+tp</u> CCMPREHT e 1994 2013 WIDE MIT, ERCIM KDD), ALL ROHTS RESERVED. WID UMBLITY, TRADEMARK, DOCUMENT US AND SITUATE MALES APPLY, YOUR INTERACTIONS WITH THES SITE ARE IN ACCOMPREHT e 1994 2013 WIDE MIT, ERCIM KDD), ALL ROHTS RESERVED. WID UMBLITY, TRADEMARK, DOCUMENT US AND SITUATE	validatic ake

Figure 2(a)-2(d) Reports Generated by W3C Validator Tool

Figure 3and 4 shows the keyword analysis and SEO score computed by employing open source SEO tools freely available on Internet.

← → C a Secure https://keywordtool.io/search/google	2/16151886?keyword=csiber&country=in&language=	en&category=web#suggestions		☆ 🖪 :
Find Keywords Analyze Competitors	Check Search Volume		Keyword Tool Pro API Access -	Login
Google YouTube Bing Amazon eBay	App Store			
All 💌 csiber'	India	English - Q		
Filter Results	Keyword Suggestions Questions		Sort by Keywords - ascending	•
Negative Keywords 🖗 🔨	Search for "csiber" found	72 unique keywords	👌 Copy all 💆 Exp	port all +
	Want to get up to 2x more keywords in	nstead? Subscribe to Keyword Tool Pro now!		
	Keywords I G	Search Volume 😡	CPC Competition	
	Csiber logo	10,000	4.00 8.00	_
	Cyber crime	Do You N	leed This Information?	
	Csiber kolhapur msw	Keyword Tool Pro For Englis	h Keywords. Try It Now!	
	Csi berkeley			
	csiber result 2017	101,000	100 3.00	
	Csiber kolhapur courses	46,200	1.01 1.00	
	Csiber result	81.000	Activate Windows Go to Settings to activat	ŝ le Windaws.
	csi bergamo	10,000	1.01 1.01	着 Copy (0)

Figure 3. Keyword Analysis Tool

Dr. P.G. Naik et al, International Journal of Advanced Research in Computer Science, 9 (1), Jan-Feb 2018,323-340

C www.checkseoreports.com/dom	in/siberindia.edu.in		Ŕ
🔪 Check SE	O Reports Home Site vs Site Recent Sites Contact	😔 EN	
Siber	india.edu.in	Fair, valid & grain valid & gr	TOEFL EARN MORE
Return to top			
Ltd Overview	>		
Q, SEO	> Sib	erindia.edu.in	
🖞 Usability	Chhatrapal Shake installate of Beatrans Education Dece	mber, 29 2017 07:34:22 PM	\frown
D Mobile	> 200 Final Action of the set of	assed	(10)
🖋 Technologies		nprove	40 Score
🖒 Social	> 03 E	rrors	
Visitors	>		
% Link Analysis	>	ownload Report C Update 🖈 Compare G Shar	Activate Windows Go to Settings to activate Window

Figure 4. Open Source Software for Computation of SEO Score

Figure 5(a)- Figure 5(d) shows the GUI of W3C validator tool implemented by the authors. Figure 5(a) depicts the ruleset currently employed by the tool. On analyzing the HTML document entered by the end user for the presence and absence of certain tags, the evaluation criteria computed by the tool is depicted in Figure 5(b). Figure 5(c) depicts in-depth analysis of HTML tag set. Based on the compliance of the rules, the score is computed and is displayed to an end user as depicted in Figure 6(a) and 6(b) for two different HTML documents.

		Netsape			
	Aucilysis Too	msn* excite	vista L		
		Inktomi	-		
S	ome Prominent SEO Rules				
S Sr. No	ome Prominent SEO Rules				
S Sr. No	Rule Rule 1, Every tag must be closed explicitly. Empty tag must use terminating slash				
Sr. No	Ome Prominent SEO Rules Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory				
Sr. No	Rule Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be in lower case				
Sr. No	Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tage must be in lower case 4.1 attributes must be enclosed in double quotes				
Sr. No	Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be in lower case 4. All attributes must be enclosed in double quotes 5. Form tags cannot be nested				
Sr. No	Ome Prominent SEO Rules Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be in lower case 4. All attributes must be enclosed in double quotes 5. Form tags cannot be nested 6. Use entity references wherever necessary				
Sr. No	Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be inclosed in double quotes 4. All attributes must be nested 6. Use entity references wherever necessary 7. CSS rules about be inclosed.				
Sr. No	Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be in lower case 4. All attributes must be enclosed in double quotes 6. Form tags cannot be nested 6. Use entity references wherever necessary 7. CSS rules should be in lower case				
Sr. No	Rule 1. Every tag must be closed explicitly. Empty tag must use terminating slash 2. The head tag is mandatory 3. All tags must be in lower case 4. All attributes must be enclosed in double quotes 5. Form tags cannot be nested 6. Use entity references wherever necessary 7. CSS rules should be in lower case				



					-
EO /	AL Page Analysis f	To To	exote overture	A AOL	
HTMI Flor	unt Total	Correct	Incorrect	ı	
Tage	17 Tota	Tage 15 Lowercase Tag	2 Lippercase Tage		
Attributes	4 Total Attr	butes 1 Quoted Attributes	3 Attributes without Quotes		
Autoutes	4 Total Att	T	5 Autobates Willout Quotes		
Start Taga	17 Total Star	1309			
Start Tags	17 Total Star 7 Total En	Tags			
Start Tage End Tags	17 Total Star 7 Total En 7 Total Terminate	Tags Tags			
Start Tags End Tags Terminated Non-Termin	17 Total Star 7 Total En Tags 0 Total Terminate ated Tag 3 Total Empt	Tags Tags Tags Tags			

Figure 5(a)-5(d) GUI of Custom W3C Validator Tool

G Form1			- ö ×
View Analyze Report	ala Too	Verture Overture O	
Evaluation Schem	e		
Rule Significant (Yes/No) 1 Yes 2 Yes 3 Yes 4 Yes 5 No 7 No	Percentage (%) Percentage Assigned(25% 25% 25% 25% 25% 25% 0% 0% 0% 0%	6) 25% 0% 25% 0% 0% 0% 0%	
SEO Compliance Scor	e 50%		
		-	
	seo	×	
	300	~	
	1 Tag(s) of 7 total t	ags do not match	
	2	-	
		OK	
Ph Form1			×
View Analyze Report			
SED Anal	ala Too	exeite overture	
Evaluation Schen	ne		
Rule Significant (Yes/No)	Percentage (%) Percentage Assigned 25% 21.428571	%) 4285714%	
2 Yes	25%	25%	
4 Yes	25%	25%	
5 No 6 No	0%	0%	
7 No	0%	0%	
SEO Compliance Sco	re 96.428571428		

Figure 6(a)-6(b) SEO Compliance Score Computed by Custom W3C Validator Tool

Figures 7(a)-(c) report the analysis of HTML document, seo2.html, with Figure 8. depicting the report generated by the tool in text format.



Dr. P.G. Naik et al, International Journal of Advanced Research in Computer Science, 9 (1), Jan-Feb 2018,323-340



Figure 8. Report Generated by the W3C Validation Tool in Text Format

Converting the Invalid HTML document in to Valid HTML Document.

The structure of HTML document, "seo.html" input to the tool is shown in Figure 9. As revealed from the HTML content head tag is missing, some tags (hr and html end tag) are in uppercase. Br tags are not terminated.



Figure 9. Structure of HTML Document

Figure 10(a) and (b) depict the report generated on performing the tag analysis by the tool in grid and text format, respectively.

View	Analyze Convert Rep-	ort					
	220	DA	alysi	5 70	Cogle Cogle Control Coverture O Coverture O Covertur		
		HTML P	age Analysis for D	:\seo.html			
		HTML Element Tags Attributes Start Tags End Tags Terminated Tags Non-Terminated Tag	Total 13 Total Tage 4 Total Attributes 13 Total Start Tage 5 Total End Tage 1 Total Terminated Tage 2 Total Non-Terminated Tage	Correct 11 Lowercase Tage 0 Quoted Attributes	Incorrect 2 Uppercase Tags 4 Attributes without Quotes		
					View		
ſ	/ sea hi	tml.txt - No	otenad				×
	File Ed	lit Forma	it View Help			 	
	*****	_ ******	******	*****	******************		*
	s	EO Comp	liance Rep	ort			
	*****	*****	******	******	******************		
	Head t	ag is m	issing				
	Total	Tags	: 13				
	Upperc	ase Tag	is : 2				
	Lowerc	ase Tag	s : 11				=
	Start	Tags :	13				
	End Ta	.gs : 5					
		Termi	nated Tags	: 1			
		Non-T	erminated	Tags :	2		
	*****	******	*******	******	******************		
	SEO CO	mplianc	e Score :	39.90%			
	*****	******	*******	******	******************		-

Figure 10(a)-(b). Report Generated by the W3C Validation Tool on Performing Tag Analysis

On selecting the, "Convert" option from the main menu, a new file with the name <filename>_new.html is created taking care of all the rules, where <filename> is the primary filename of the HTML file entered by an end user. The structure of the new document created by the tool is shown in Figure 11.



Dr. P.G. Naik et al, International Journal of Advanced Research in Computer Science, 9 (1), Jan-Feb 2018, 323-340



Figure 11. Structure of the Valid HTML Document Generated by the Tool

On analyzing the converted HTML document, the reports generated by the tool are depicted in Figure 12.







Figure 12. Reports Generated on Analyzing HTML Document Generated by Tool

5. CONCLUSIONS AND SCOPE FOR FUTURE WORK

Search engine optimization conforms to the act of optimizing the website so as to acquire a highest ranking in search engine results. The ranking is based on some predefined criteria which SEO encapsulates for generating right signal to the search engine. Hence knowledge of these criteria enables in rendering the website visible to most of the search engines and being listed at the top of the search results generated by a search engine. In the current research, the authors have considered a set of validation rules for W3C compliance and a tool is designed and implemented to generate a W3C Compliance score for a web page. The tool also enables converting an invalid HTML document to its valid counterpart by incorporating the necessary changes to the document by referring to the rule set. The tool acts as a first-hand simulation tool to understand search engine optimization issues. The scope of the tool can be further broadened to incorporate few more rules corresponding to page ranking, back linking, keyword density calculation etc. to widen its scope.

REFERENCES

 Constantinos Dimopoulos, Sergey Nepomnyachiy, TorstenSuel, "Optimizing Top-k Document Retrieval Strategies for Block-Max Indexes", IEEE Conference, ACM, 2013, 113-122.

- Lihong Zhang, Jianwei Zhang, YanbinJu, in "The Research on Search Engine Optimization Based on Six Sigma Management", in IEEE conference 2011, 101-103
- 3. Antriksha Somani, Ugrasen Suman , "Counter Measures against Evolving Search Engine Spamming Techniques", IEEE Conference, 2011 , 214-217.
- H. L. Hsing, H. J. Chen, L. Me. Huang and H. H. Yi, "The application of search engine optimization for internet marketing: An example of the motel web sites", ICCAE Chung Hua University Taiwan, vol. 1, (2010), pp. 380-383.
- P. T. Chung, S. H. Chung and C. K. Hui, "A web server design using search engine optimization techniques for web intelligence for small organizations", LISAT IEEE Long Island University Brooklyn NY USA, (2012), pp. 1-6.
- M. Cui and S. Hu, "Search Engine Optimization Research for Website promotion", ICM Transp. Manage., China, vol. 4, (2011), pp. 100-103.
- Killoran, J.B., "How to Use Search Engine Optimization Techniques to Increase Website Visibility", Professional Communication, IEEE Transactions on (Volume:56, Issue: 1)
- S. M. Kumar, P. Revathy and K. Vijayalakshmi, "Implementation of two-tier link extractor in optimized search engine filtering system", IEEE IMSAA IT Dept. Hyderabad, India, (2009), pp. 1-4.
- S. Wang, K. Xu, Y. Zhang and F. Lie, "Search Engine Optimization based on Algorithm of BP Neural networks", CIS Inst. of Eng. & Tech., China, (2011), pp. 390-394
- M. Rajaram and S. L. S. Vadivu, "Web caching in Semantic Web based multiple search engines", ICCIC IEEE Dept. of EEE Tirunelveli, India, (2010), pp.1-7.

Appendix A Partial Source Code for Implementation of SEO Compliance Test Tool in Visual Basic 6

head = False form1 = False css = False Tagupper = False uppercnt = 0 cattr = 0 total = 0 mnuEvaluation_Click If (head = True) Then total = total + percentage End If Open filename For Input As #1 While (EOF(1) = False)

/hile (EOF(1) = False) Input #1, Line1 n = InStr(1, Line1, "<") m = InStr(1, Line1, ">")

 $\begin{array}{l} \mbox{While } (n > 0 \mbox{ And } m > 0) \\ \mbox{ Tag = Mid(Line1, n + 1, m - 2) } \\ \mbox{ If } (Tag = UCase(Tag)) \mbox{ Then } \\ \mbox{ Tagupper = True } \\ \mbox{ uppercnt = uppercnt + 1 } \\ \mbox{ End If } \\ \mbox{ n = InStr(n + 1, Line1, "<") } \\ \mbox{ m = InStr(m + 1, Line1, ">") } \end{array}$

Wend Wend Close #1 Open filename For Input As #1 While (EOF(1) = False)Input #1, Line1 n = InStr(1, Line1, "=")While (n > 0)n = InStr(n + 1, Line1, "=")tattr = tattr + 1Wend Wend Close #1 Open filename For Input As #1 While (EOF(1) = False)Input #1, Line1 n = InStr(1, Line1, "=")m = Mid(Line1, n + 1, 1)While (n > 0 And m = """")Wendattr = Truen = InStr(n + 1, Line1, "=")m = Mid(Line1, n + 1, 1)cattr = cattr + 1Wend Close #1 score2 = cattr / tattr * percentage

count = 4

Open filename For Input As #1 While (EOF(1) = False)Input #1. Line1 n = InStr(1, "<head>", Line1)If (n > 0) Then head = TrueEnd If n = InStr(1, "form", Line1)If (n > 0) Then form1 = True count = count + 1End If n = InStr(1, "style", Line1)If (n > 0) Then css = Truecount = count + 1End If n = InStr(1, "&", Line1)If (n > 0) Then entity = Truecount = count + 1End If Wend Close #1 percentage = 100 / countn = countstarttagsm = countempty endtags $\mathbf{x} = \mathbf{n} - \mathbf{m}$ y = countendtagslowercnt = n - uppercnttotalcnt = nscore1 = lowercnt / totalcnt * percentage If (x = 2 * y) Then MsgBox "Tags match.." Else MsgBox Abs(x - 2 * y) &" Tag(s) of " & (x - y + m) & " total tags do not match..' End If diff = (x - y + m) - Abs(x - 2 * y)score = diff / (x - y + m) * percentage If (head = True) Then MSFlexGrid1.TextMatrix(2, 3) = Format(CStr(percentage), "##.00") & "%" Else MSFlexGrid1.TextMatrix(2, 3) = "00.00%" End If If (Tagupper = True) Then MSFlexGrid1.TextMatrix(3, 3) = Format(CStr(score1), "##.00") & "%"

Else MSFlexGrid1.TextMatrix(3, 3) = Format(CStr(percentage), "##.00") & "%" End If End Sub Private Sub mnuConvert Click() Dim Line1 As String fname = Left(filename, Len(filename) - 5)fname = fname& "_new.html" Open fname For Output As #2 Open filename For Input As #3 While (EOF(3) = False)Input #3, Line1 n = InStr(1, Line1, "=")If (Trim(LCase(Line1)) = "<html>") Then Print #2, Line1 If (head = False) Then Print #2, "<head>" Print #2, vbTab& "<title>" & filename & "</title>" Print #2, "</head>" End If ElseIf (Trim(LCase(Line1)) = "<hr>") Then Print #2, "<hr/>" ElseIf (Trim(LCase(Line1)) = "
") Then Print #2, "
 ElseIf (n > 0) Then m = Mid(Line1, n + 1, 1)While (n > 0 And m > 0)If (m <> """") Then s1 = InStr(n, Line1, "") s2 = InStr(n, Line1, ">")If (s1 > s2) Then Min = s2Else Min = s1End If newline = Mid(Line1, 1, n) & """" & Mid(Line1, n + 1, (Min - n - 1)) & """" & Mid(Line1, (Min), Len(Line1) - Min + 1)Line1 = newlineEnd If n = InStr(n + 1, Line1, "=")m = Mid(Line1, n + 1, 1)Wend Print #2, LCase(Line1) Else Print #2, LCase(Line1) End If Wend Close #3 Close #2 MsgBox "Converted Successfully and Saved As "&fname Shell ("notepad "&fname)

Private Sub mnuEvaluation_Click() Dim count As Integer

End Sub

Dim Line1 As String

Label1.Visible = False MSFlexGrid1.Visible = False

CommonDialog1.Action = 1filename = CommonDialog1.filename count = 4Open filename For Input As #1

While (EOF(1) = False)Input #1, Line1 n = InStr(1, "<head>", Line1)

If (n > 0) Then head = TrueEnd If

n = InStr(1, "form", Line1)If (n > 0) Then form1 = Truecount = count + 1End If

n = InStr(1, "style", Line1) If (n > 0) Then css = Truecount = count + 1End If

n = InStr(1, "&", Line1)If (n > 0) Then entity = True count = count + 1End If Wend Close #1 percentage = 100 / count =End If End If End If End Sub

Private Sub mnuSeoCompliance Click() Label1.Caption = "HTML Page Analysis for "& filename

uppercnt = totalcnt - lowercnt wattr = tattr - cattr

n = countempty endtagst = nn = countstarttags - 2 * countendtags - countemptyendtags

Open filename & ".txt" For Output As #2

Print #2. *****" &vbCrLf Print #2, " SEO Compliance Report" &vbCrLf

Print

#2. ******" &vbCrLf&vbCrLf If (head = False) Then Print #2, "Head tag is missing" &vbCrLf End If Print #2, "Total Tags : " &totalcnt&vbCrLf Print #2, "Uppercase Tags : " &uppercnt&vbCrLf Print #2, "Lowercase Tags : " &lowercnt&vbCrLf Print #2, "Start Tags : " &countstarttags&vbCrLf Print #2, "End Tags : " &countendtags&vbCrLf Print #2, vbTab& "Terminated Tags : " & t &vbCrLf Print #2, vbTab& "Non-Terminated Tags : " & n &vbCrLf Print #2. *****" &vbCrLf Print #2, "SEO Compliance Score : " & Label3.Caption &vbCrLf Print #2 *****" &vbCrLf Close #2 MsgBox "Report saved as "& filename & ".txt" End Sub Public Function countstarttags() As Integer Dim Line1 As String Dim cnt As Integer cnt = 0Close #1 Open filename For Input As #1 While EOF(1) = FalseInput #1, Line1 n = InStr(1, Line1, "<")While (n > 0)cnt = cnt + 1n = InStr(n + 1, Line1, "<")Wend Wend Close #1 countstarttags = cnt End Function Public Function countemptyendtags() As Integer Dim Line1 As String Dim cnt As Integer Open filename For Input As #1 cnt = 0While EOF(1) = FalseInput #1, Line1 m = InStr(1, Line1, "/>")While (m > 0)cnt = cnt + 1m = InStr(m + 1, Line1, "/>")Wend Wend Close #1 countemptyendtags = cnt End Function Public Function countendtags() As Integer Dim Line1 As String Dim cnt As Integer

cnt = 0

Open filename For Input As #1 While EOF(1) = False Input #1, Line1 n = InStr(1, Line1, "</") While (n > 0) cnt = cnt + 1 n = InStr(n + 1, Line1, "</") Wend Close #1 countendtags = cnt End Function