

Web Tags Formatting with Multilevel Numbering

Andrea Stevens Karnyoto, Marius Limpo

Department of Information System, Engineering Faculty, Toraja Indonesia Christian University

Keywords:

HTML tags
Website characteristic
Tags mapping

ABSTRACT

To know the headline location in a web page/blog automatically needed an extraction method because each website/blog has own unique characteristic in advertisement, headline, and link list placement. This research developed a system to mapping tag pairs, and single tag of HTML on website/blog code, the result can be used as a pattern or comparison against already saved patterns in the database. Extracting and mapping method use both vertical and horizontal multilevel numbering. Data example has taken from 20 websites of government, school, and also private companies to obtain the result of tag mapping that is going to be used to know the amount of pattern that produced by a website. Template comparisons are up to 5th level, the result of extracted website/blog contains hundreds of pages starting from the lowest that is 110 and the highest 280 and it is only less than ten patterns for each website. Conclusion of this research is that all extracted website has pattern less than 10% of the total pages.

*Copyright © 2013 Information Systems International Conference.
All rights reserved.*

Corresponding Author:

Andrea Stevens Karnyoto
Department of Information System, Engineering Faculty,
Toraja Indonesia Christian University,
Jl. Nusantara No. 13, Makale, Tana Toraja, South Sulawesi 90000, Indonesia.
Email: karnyoto@gmail.com

1. INTRODUCTION

Over the past decades, researches in computing science and information technology have been intensively carried out worldwide to pervade computer systems into every area of human life [1]. Managing knowledge on the World Wide Web has become an important issue to given the large volume of information that is now available on the Internet. However, the management of this knowledge is a difficult task both because of the dynamic nature of the Internet [2].

Currently, most Web content written in HTML, which follows a rigid format in displaying content because web pages for the syntax based HTML Web written for human comprehension [2]. The HEAD element contains information, also called metadata, about a web page, such as its title, keywords, description, language, and place of publication that may be useful to search engines, and other metadata that not considered page content [3]. The purpose of Meta tags are search engines can find key information from the website [4], [5]. The Meta tag provides authors and Web site owners a means to control how their information displayed and retrieved in a search engine [6]. Observed that collaborative tagging users exhibit a great variety in their sets of tags; some users have many tags, and others have few [7]. Other research has revealed from 27% to 38% of website pages containing Meta tags descriptions [8]. It means that 62% up to 73% of web pages do not have Meta tags descriptions.

Meta tags description has characters length limitation [9], [10]. There are many web pages which do not have Meta tags description [11], Meta tags description are not suitable with the web page content and Meta tags description only describes website thoroughly, not per page. Creating a pattern must tags based because tags have standard rules in use and BODY element of website consists of tags, news and links. Therefore, author created a method to extract and map the BODY element on every website page so that each tag converts to unique series number in a tiered series numbering called multilevel numbering method. A website page converts to a tiered series numbering, if the tiered series numbering does not exist in the database then the tiered series numbering going to stored as a pattern on the database in others it saves as a comparison result. Author created a system using multilevel numbering method to extract and map the

information in BODY element based on web tags. The result used as a pattern or comparison material against the pattern. It will be useful such as to know the main content of a website page.

2. RESEARCH METHOD

Purpose of this research created an extraction method called multilevel numbering, building a system based on the method, test the system against 20 websites that have been online, count total web tags pattern and error page presentation. System features are able to perform the reading three website pages at once, process of patterns extraction and comparison to stored pattern in the database is running as background.

The use of tags to identify potential description wording usually text of Web pages had proved relatively unsuccessful [8]. Multilevel numbering is a level that contains level groups put into order based on the number according to position of the element. Mapping a web page can be a pattern or comparison material to patterns used Tags in the BODY element of web code.

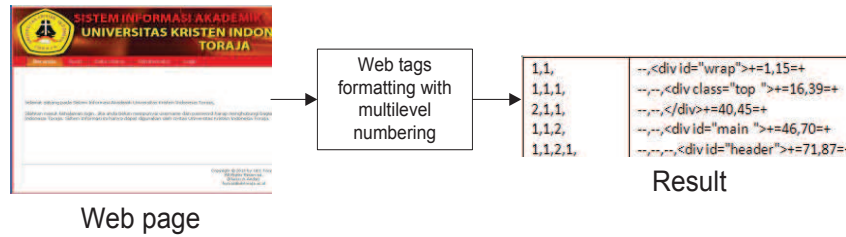


Figure 1. Overview of system

In general, system extracts web page code into level numbering arrangement (see figure 1). The developed system takes all pages from a website and processes it one by one. Thus, each page has an extraction result.

Every page on a website going to extracted to know the level of tags in it. The result going to compared with stored patterns in the database. If a pattern is not the same with the patterns in the previous pages, the pattern saves as a new pattern in the database. This system has three classes, *TDownloadPageHttp*, *TPartingStringHtml* and *TDatabaseAdder*.

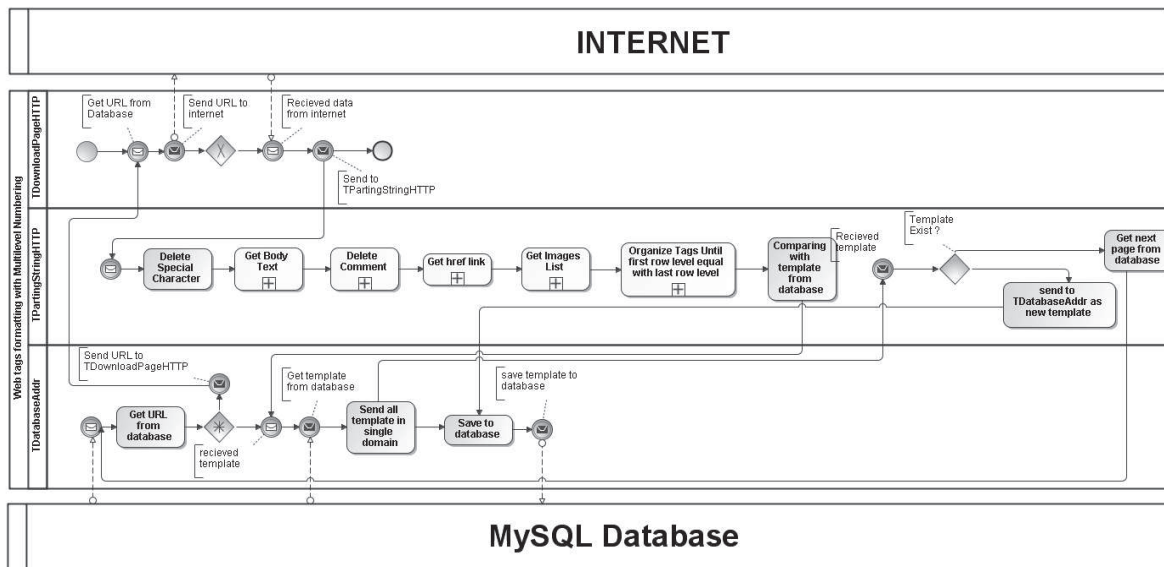


Figure 2. Collaborative Process

TDownloadPageHttp class function is to take website page from the internet and send it to *TPartingStringHtml*. *TPartingStringHtml* function is to conduct extraction and mapping to the page has downloaded. *TDatabaseAdder* class function is to keep extraction result in the database if the extraction result does not exist in database (See Figure 2).

Part of downloaded code page only takes from BODY element. Tags inside the BODY element going to extracted to multilevel numbering mode.

1,1,2,2,1,2,1,4,2,	--,-,-,-,-,-,-,-, +=2529,2534=+				
2,1,2,2,1,2,1,4,2,	--,-,-,-,-,-,-,-, +=2559,2565=+				
<table style="display: inline-table; border: none; text-align: center;"> <tr><td style="padding: 0 5px;">A</td><td style="padding: 0 5px;">B</td></tr> </table> <table style="display: inline-table; border: none; text-align: center;"> <tr><td style="padding: 0 5px;">C</td><td style="padding: 0 5px;">D</td></tr> </table>	A	B	C	D	
A	B				
C	D				

Figure 3. Web Tags Map Formatting

Figure 3 shows A = number 1 is for opening and number 2 is for closing of tag, B = level of tag, C = tag in the web page, D = position of tag in a web page.

Table 1. Tags Position on Multilevel Numbering

A	B(1)	B(2)	B(3)	B(4)	B(5)	B(6)	B(7)	B(8)	B(n)
1	1	2	2	1	2	1	4		
1	1	2	2	1	2	1	4	1	
2	1	2	2	1	2	1	4	1	
1	1	2	2	1	2	1	4	2	
2	1	2	2	1	2	1	4	2	
2	1	2	2	1	2	1	4		

See Table 1, the opening and closing tag which are partners, have the same level, but they have different code in column A, number 1 is for opening code and number 2 is for closing the code. Meanwhile, B (1) up to B (n) is the position of tags in multilevel numbering order.

1,1,2,1,2,	--,-,-,-,-,-, <div id="menu">+=188,202=+
1,1,2,1,2,1,	--,-,-,-,-,-,-, +=203,206=+
1,1,2,1,2,1,1,	--,-,-,-,-,-,-,-, +=207,210=+
2,1,2,1,2,1,1,	--,-,-,-,-,-,-,-, +=242,246=+
1,1,2,1,2,1,2,	--,-,-,-,-,-,-,-, +=247,250=+
2,1,2,1,2,1,2,	--,-,-,-,-,-,-,-, +=311,315=+
1,1,2,1,2,1,3,	--,-,-,-,-,-,-,-, +=316,319=+
2,1,2,1,2,1,3,	--,-,-,-,-,-,-,-, +=369,373=+
1,1,2,1,2,1,4,	--,-,-,-,-,-,-,-, +=374,377=+
2,1,2,1,2,1,4,	--,-,-,-,-,-,-,-, +=433,437=+
1,1,2,1,2,1,5,	--,-,-,-,-,-,-,-, +=438,441=+
2,1,2,1,2,1,5,	--,-,-,-,-,-,-,-, +=497,501=+
2,1,2,1,2,1,	--,-,-,-,-,-,-, +=557,561=+
2,1,2,1,2,	--,-,-,-,-,-, </div>+=562,567=+

Figure 4. Web Tags Level Trim

The method to take first tag up to the fifth level is by erasing all tags located beyond fifth level as it shown in figure 4, the shaded tags erased automatically so that the picture become as shown in figure 5.

1,1,2,1,2,	--,-,-,-,-,-, <div id="menu">+=188,202=+
1,1,2,1,2,1,	--,-,-,-,-,-,-, +=203,206=+
2,1,2,1,2,1,	--,-,-,-,-,-,-, +=557,561=+
2,1,2,1,2,	--,-,-,-,-,-, </div>+=562,567=+

Figure 5. After Trim Tag Result

After trim tags result is a pattern as shown on figure 5, it is the final result from extraction and mapping process of a website page. The pattern is going to save in the database and pattern contains about level information such as the number from tag position, sorted tag, start and end position of tag on a web page source code. The pattern also used for comparison material to other web page patterns in a website.

3. RESULTS AND ANALYSIS

Sample data to test the constructed system taken from 20 websites, including website of government, school, and private company. Each tag extraction results going to use to know how many patterns resulted from a website, the comparison of the template is up to 5th level of numbering.

System uses three threads for parallel download, extract, and save pattern to database (see figure 6). Data of extracted pages can be seen at figure 7. All templates and related information of the page saved in the database.

Page Name	Thread	Byte	
http://www.upxanel.com	2	13010	Done
http://ukitoraja.ac.id	3	17883	Done
http://resensifilmbagus.blogspot.com	3		Start
http://www.pa-amurang.go.id	2	38390	Done
http://www.pa-sidrap.go.id	2	71462	Done
http://karnyoto.blogspot.com	3		Start
http://www.sman1-pacitan.sch.id	2	51534	Done
http://www.sekolahciputra.sch.id	2		Start
http://www.acsjakarta.sch.id			
http://www.sekolahbogorraya.com			
http://www.amartakarya.co.id			
http://www.anneahira.com			
http://www.metrodata.co.id			
http://idrusrudeng.wordpress.com			
http://www.akprind.ac.id			
http://dprd-mamujuutarakab.go.id			
http://dprd-pareparekota.go.id			

Figure 6. The Process of Page Extracting

Nama Halaman	Tmp	Title	Tanggal
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=36&Itemid=8	2	proses beracara	6/4/2013 1:20:25 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=26&Itemid=7	2	pendaftaran perkara	6/4/2013 1:20:30 PM
http://www.pa-amurang.go.id/	2	situs resmi pengadilan	6/4/2013 1:20:31 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=69&Itemid=77	2		6/4/2013 1:20:28 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=67&Itemid=76	2	panduan pengajuan	6/4/2013 1:20:32 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=68&Itemid=75	2		6/4/2013 1:20:33 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=124&Itemid=7	2	buku panduan	6/4/2013 1:20:42 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=frontpage&Itemid=27	2	situs resmi pengadilan	6/4/2013 1:20:44 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=56&Itemid=67	2	statistik pengadilan	6/4/2013 1:20:43 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=55&Itemid=66	2	data hukuman	6/4/2013 1:20:48 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=53&Itemid=64	2	mekanisme pengaduan	6/4/2013 1:20:51 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=52&Itemid=63	2	pedoman pengaduan	6/4/2013 1:20:50 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=50&Itemid=59	2	hak-hak pencari keadilan	6/4/2013 1:20:54 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=frontpage&Itemid=25	2	situs resmi pengadilan	6/4/2013 1:20:56 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=89&Itemid=94	2		6/4/2013 1:20:52 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=87&Itemid=93	2	dipa 2012	6/4/2013 1:20:58 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=90&Itemid=95	2		6/4/2013 1:20:57 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=article&id=88&Itemid=92	2		6/4/2013 1:20:59 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=category&id=7&Itemid=8	2	transparansi	6/4/2013 1:21:01 PM
http://www.pa-amurang.go.id/index.php?option=com_content&view=category&id=29&Itemid=	2	laporan uang iwad	6/4/2013 1:21:03 PM

Figure 7. List of page download

Table 2. The Result of Extracting Websites

Website	Pages	Templates	Errors
http://www.upxanel.com	129	4	0
http://ukitoraja.ac.id	110	3	1
http://resensifilmbagus.blogspot.com	280	6	0
http://www.pa-amurang.go.id	214	9	2
http://www.pa-sidrap.go.id	271	7	3
http://karnyoto.blogspot.com	153	4	0
http://www.sman1-pacitan.sch.id	122	6	3
http://www.sekolahciputra.sch.id	112	4	2
http://www.acsjakarta.sch.id	206	8	6
http://www.sekolahbogorraya.com	168	3	4
http://www.amartakarya.co.id	139	2	0
http://www.anneahira.com	113	4	0
http://www.metrodata.co.id	114	4	0
http://idrusrudeng.wordpress.com	117	3	0
http://www.akprind.ac.id	167	7	3
http://dprd-mamujuutarakab.go.id	123	4	4
http://dprd-pareparekota.go.id	140	6	1
http://www.barrukab.go.id	181	8	3
http://sementonasa.co.id	111	3	3
http://www.ekaristi.org	194	4	2

Table 2 shows system execution results to perform the extraction and mapping of pages to 20 websites. Table columns shows the amount of processed pages, amount of template result by using multilevel numbering, and amount of pages cannot be processed. Error while conducting a mapping and extraction on several pages of website caused by incomplete tag syntaxes, it written by website programmer even though the browser can still read or ignore it.

4. CONCLUSION

Contents of the BODY element are a combination of news and web tags, a method required to extract and map web tag into unique series number, and this research uses multilevel numbering method.

All extracted websites have a pattern is less than 10% that number of whole pages in a website. The least amount of the page is 110, and the highest is 280 pages. The system produces less than 3.5% error by using multilevel numbering method. The system result is web tag pattern and it are can be material to determine the position of headlines, advertisements and links list. All components position listed in unique tiered series numbering. Sometimes, the web programmer was not paying attention in writing complete web tags of a website. Therefore, there were tags which had an opening but not a closing element or otherwise. Thus, the system reads as “error”. It suggested doing further research to get the news of a website by using this multilevel numbering method, and researches formatting website tags by using other methods.


ACKNOWLEDGEMENTS

This research was fully funded and supported by Department of Information System, Engineering Faculty, Toraja Indonesia Christian University.

REFERENCES

- [1] Y. Bassil and M. Alwani, "Autonomic Html Interface Generator For Web Applications", *International Journal of Web & Semantic Technology*, vol. 3(1), pp. 1-15, January 2012.
- [2] T.C. Du, *et al.*, "Managing knowledge on the Web – Extracting ontology from HTML Web", *Decision Support Systems*, vol. 47(4), pp. 319–331, Nov 2009.
- [3] A. Noruzi, “A Study of HTML Title Tag Creation Behavior of Academic Web Sites”, *Journal of Academic Librarianship*, vol. 33 (4), pp. 501–506, 2007.
- [4] H. Joseph Wen, *et al.*, “E-commerce Website design:strategies and models”, *Information Management & Computer Security*, vol. 9(1), pp. 5-12, 2001.
- [5] H. Han and R. Emasri, “Learning Rules for Conceptual Structure on the Web”, *Journal of Intelligent Information Systems*, vol. 22(3), pp. 237-256, 2004.
- [6] R. Henshaw," The First Monday Metadata Project", *Libri*, 1999, pp. 125-131.
- [7] S.A. Golder and B.A. Huberman, “Usage Patterns of Collaborative Tagging Systems”, *Journal of Information Science*, vol. 32(2), pp. 198–208, 2006.
- [8] T. C. Craven, *et al.*, “HTML Tags as Extraction Cues for Web Page Description Construction”, *Informing Science Journal*, Vol. 6, pp. 1-12, 2003.
- [9] A. Riad, *et al.*, “Web Image Retrieval Search Engine based on Semantically Shared Annotation”, *IJCSI International Journal of Computer Science*, vol. 9, Issue 2, No 3, March 2012.
- [10] A.W. Huang and T.R. Chuang, “Social Tagging, Online Communication, and Peircean Semiotics: A Conceptual Framework”, *Journal of Information Science*, vol. 20(10), pp. 1-18, 2008.
- [11] S. Gupta, *et al.*, “Automating Content Extraction of HTML Documents”, *World Wide Web Journal*, vol. 1, pp. 3-7, 2004.

BIBLIOGRAPHY OF AUTHORS

	<p>Andrea Stevens Karnyoto, S.Kom.,MT., born in Sept 8th, 1979, obtained his Bachelor in Informatics Technology from the STMIK Dipanegara, Makassar, Indonesia, in 2003, his Master in Electrical Engineering with Informatics Technology concentrate (2010) from the Hasanuddin University. He has been head of the informatics technologies consultant in CV. Anugrah Empat Pilar starting from 2007, and lecture in Toraja Indonesia Christian University of the department Informatics Technology starting from 2012.</p>



Marius Limpo, S.Kom., born in Makassar, may 19th, 1985, obtained his Bachelor in informatics Technology from ATMA JAYA University Makassar, Indonesia 2010. Researchers Student at KYUSHU UNIVERSITY Japan, from October 2011 until March 2012 and now lecture in Toraja Indonesia Christian University of the department informatics Technology starting from 2012.