HTML Tags Formatting with Tree Node

Andrea Stevens Karnyoto and Marius Limpo
Department of Information System, Engineering Faculty
Toraja Indonesia Christian University
Makale, Tana Toraja, South Sulawesi, Indonesia
karnyoto@gmail.com, mariulimpo@ukitoraja.ac.id

***Abstract*— a method of HTML tags mapping is needed in order to know the position of headlines of website/blog page automatically and immediately because every website/blog has its own characteristics in placing the advertisement, headline, and list of link. A system for mapping HTML tags on website/blog code is needed. The result will be used as pattern and comparison from a processed pattern which had been saved. Mapping method uses multilevel node whereas each node has its parents and each parent node has many child nodes. The example of data will be taken from 15 websites, including government, school, and private company's website to obtain each tag mapping which are going to be used to recognize amount of patterns produced by a website, comparison of patterns up to the fifth level. The result of processing obtained from an extracted website/blog has tens of pages, the least page is 18 and the highest is 280 and only contain less than 10 patterns.**

I. INTRODUCTION

Currently, most Web content is written in HTML, which follows a rigid format in displaying content because web pages for the syntax based HTML Web are written for human comprehension [5]. 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 are not considered page content [1]. Aims Meta tags are search engines can find information key from the website. The Meta tag provides authors and Web site owners a means to control how their information is displayed and retrieved in a search engine [2]. Observed that collaborative tagging users exhibit a great variety in their sets of tags; some users have many tags, and others have few [3]. Other research has revealed from 27% to 38% of Web pages containing Meta tag descriptions [4]. It means that 62% up to 73% web page does not have Meta tags descriptions.

Because Meta tags description has characters length limited, there are many web pages which do not have Meta tags description, Meta tags description is not suitable with the web page content and Meta tags description describes only website thoroughly, not per page. Make patterns must tags based because tags have the standard rules in use and BODY element consists of tags, news, and links. Therefore, author created a method to extract and mapping BODY element in every website page so that each tag converts to unique node in a tree node called tree node method. A website page converts to a tree node; if the tree node does not have in common with the existing tree node in the database then the tree node will be stored as a pattern in others will be saved as a comparison results. Author create a system using tree node method that is able to extract and map the information in BODY element based on web tags which can be used later as pattern or comparison material against the useful pattern such as to know the main content of a website page.

II. RESEARCH METHOD

This research has purpose to create an extraction method called tree node method, building a system based on the method, test the system against 15 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 with stored pattern in database is running as background but system can show the process to user such as downloaded byte, website page name being downloaded, and download status.

The use of tags to identify potential description wording usually text of Web pages had proved relatively unsuccessful [8]. To format HTML tags to Tree Node is to change a pair tag into a node, pair tag <BODY></BODY> is root node, nodes child of root node is tags which are in the one level in it, and so it goes until the child node can be parent for node in it. Each node must have parent except parent from root node is NULL. Node does not have to have the child node.

A. General Overview

All kind sample single and paired website tags registered in database so the system can detect error writing tags by comparing target website tag to registered tags list and system will be rejected all unrecognized tags; open and close tag required to make a node from paired tag but node from single tag only open tag needed.

System recognize website tags only by sample tags has registered in database but will ignore element inside website tag, i.e. system be able recognize "<TABLE" but "WIDHT='100%'" ignored.

In general, system extracts web page into tree node level, see source code sample at fig. 1 and the result at fig. 2. System which is developed to take all pages from a website and process it one by one, is built so that each page has extraction result.

ISBN: 978-602-8509-19-0

```
<body>
  <div class="header">
     <div class="logo"></div>
     <div class="title"></div>
     <table>
       <tr>
         <td>
         </td>
         <td>
         </td>
       </tr>
     </table>
  </div>
  <div class="continer">
     <div class="breadcumb"></div>
     <div class="content">
        <div class="artikel">
           <table>
             <tr>
               <td>
               </td>
               <td>
               </td>
             </tr>
             <tr>
               <td>
               </td>
               <td>
               </td>
             </tr>
           </table>
        </div>
        <div class="artikel">
        </div>
     </div>
  </div>
</body>
```

Figure 1. HTML Source Code

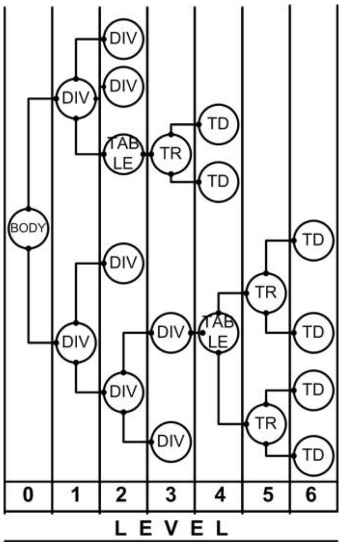


Figure 2. Result of Tree Node

B. Node Declaration

Fig. 3 shows node declaration by using Delphi 2010, all variables written in private area so only can be read from outside using encapsulation method.

```
TTagNode = class(TObject)
   private
     FTagOpen, FTagClose  : string;
     FTagLevel : Integer;
     FStartPosTagOpen, FEndPosOpen : Integer;
     FStartPosTagClose, FEndPosTagClose : Integer;
     FParent : TObject;
     FListChild : TObjectList;
   Public
      //
end;
```

Figure 3. *TTagNode* Class Declaration

C. Main Process Class

The system has three main process classes; each class has a different function. *THttpDownloadPage* class function is to take a website page source from internet and send to next process class. After getting the page source from *THttpDownloadPage* class, which *TPartingtoThreeNode* class continue the process to conduct extraction and mapping to web tags that inside BODY element. *TPartingtoThreeNode* class can detect syntax error and will not process incomplete paired tags, the website page contain incomplete paired tags or syntax error is defined as error page. *TAddtoDatabase* class is to conduct save all data processed from *TPartingtoThreeNode* class; *TAddtoDatabase* class also is duty compare web page pattern against comparison material that saved in database (see fig. 4).

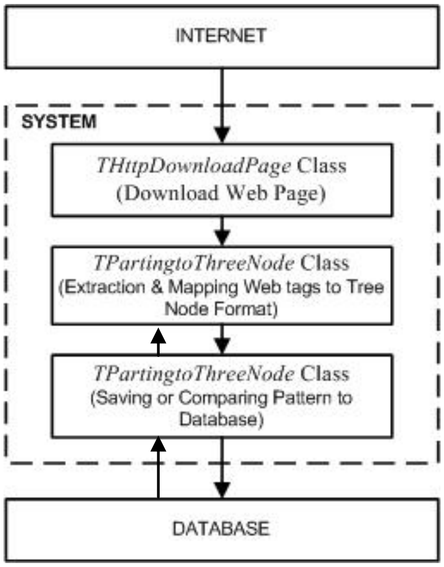

INTERNET
SYSTEM
THttpDownloadPage Class
(Download Web Page)
TPartingtoThreeNode Class
(Extraction & Mapping Web tags to Tree Node Format)
TPartingtoThreeNode Class
(Saving or Comparing Pattern to Database)
DATABASE

Figure 4. Main Process Class Flow

D. Pattern Rules

Website page pattern make by comparing up to fifth level (see fig. 2), sixth level and next are ignored. If pattern is not available in the database, pattern be saved into database and if

ISBN: 978-602-8509-19-0

is already exist be ignored. One website can have many patterns and patterns can only be used for a website. The pattern saves as RAW/BLOB type in database pattern field.

E. Data Sample

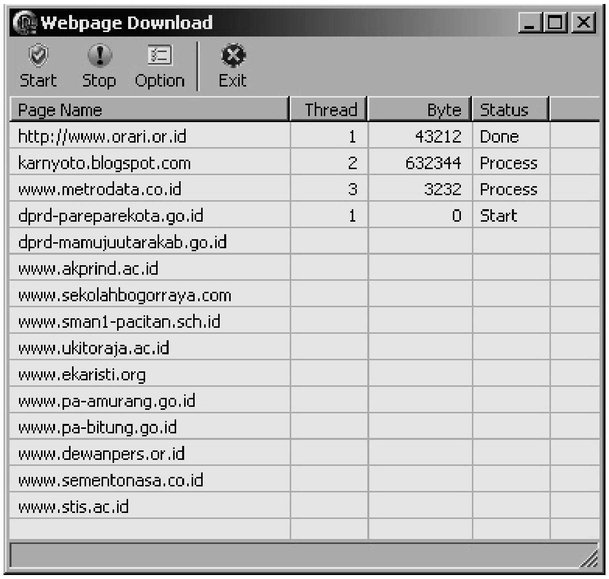
Data sample to test the system which was constructed is 15 website including government, school, and private company to obtain each result of tag extraction which will be used to know the amount of patterns produced by a website.

F. Experimental Configuration

The system is running on personal computer and connected to internet by using ASDL2 network. The computer is using windows 7 and installed MySQL 5.5 for database engine.

III. Result and Analisys

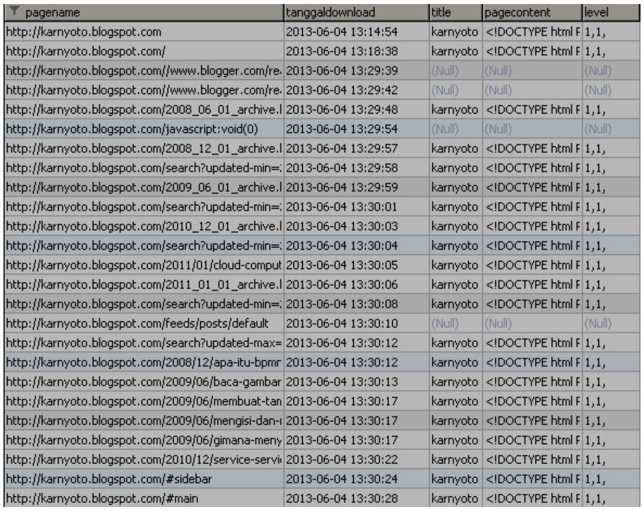
This section described our experimental result. Fig 5 shows website pages download, fig 6 shows the result of page in database table, fig 7 shows one of node saving text to database and using XML mode.



Page Name	Thread	Byte	Status
http://www.orari.or.id	1	43212	Done
karnyoto.blogspot.com	2	632344	Process
www.metrodata.co.id	3	3232	Process
dprd-pareparekota.go.id	1	0	Start
dprd-mamujuutarakab.go.id			
www.akprind.ac.id			
www.sekolahbogorraya.com			
www.sman1-pacitan.sch.id			
www.ukitoraja.ac.id			
www.ekaristi.org			
www.pa-amurang.go.id			
www.pa-bitung.go.id			
www.dewanpers.or.id			
www.sementonasa.co.id			
www.stis.ac.id			

Figure 5. Webpage Download Utility

Website pages download using three threads, while one thread is running other threads are also running until all of task completed (see fig. 5).



pagename	tanggaldownload	title	pagecontent	level
http://karnyoto.blogspot.com	2013-06-04 13:14:54	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/	2013-06-04 13:18:38	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com//www.blogger.com/re	2013-06-04 13:29:39	(Null)	(Null)	(Null)
http://karnyoto.blogspot.com//www.blogger.com/re	2013-06-04 13:29:42	(Null)	(Null)	(Null)
http://karnyoto.blogspot.com/2008_06_01_archive.l	2013-06-04 13:29:48	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/javascript:void(0)	2013-06-04 13:29:54	(Null)	(Null)	(Null)
http://karnyoto.blogspot.com/2008_12_01_archive.l	2013-06-04 13:29:57	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/search?updated-min=:	2013-06-04 13:29:58	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2009_06_01_archive.l	2013-06-04 13:29:59	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/search?updated-min=:	2013-06-04 13:30:01	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2010_12_01_archive.l	2013-06-04 13:30:03	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/search?updated-min=:	2013-06-04 13:30:04	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2011/01/cloud-comput	2013-06-04 13:30:05	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2011_01_01_archive.l	2013-06-04 13:30:06	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/search?updated-min=:	2013-06-04 13:30:08	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/feeds/posts/default	2013-06-04 13:30:10	(Null)	(Null)	(Null)
http://karnyoto.blogspot.com/search?updated-max=	2013-06-04 13:30:12	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2008/12/apa-itu-bpmr	2013-06-04 13:30:12	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2009/06/baca-gambar	2013-06-04 13:30:13	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2009/06/membuat-tan	2013-06-04 13:30:17	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2009/06/mengisi-dan-i	2013-06-04 13:30:17	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2009/06/gimana-meny	2013-06-04 13:30:17	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/2010/12/service-servi	2013-06-04 13:30:22	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/#sidebar	2013-06-04 13:30:24	karnyoto	<!DOCTYPE html F	1,1,
http://karnyoto.blogspot.com/#main	2013-06-04 13:30:28	karnyoto	<!DOCTYPE html F	1,1,

Figure 6. Result of Page Download in Database Table

Fig 6 shows complete data of website pages information, every website pages record saved information field (page name, date of download, title page, page content, level, domain, tags result, etc).

```
......
<node number=214>
  <FTagOpen>~div class=”content”~</FTagOpen>
  <FTagClose>~div~</FTagClose>
  <FTagLevel>4</ FTagLevel>
  <FStartPosTagOpen>321</FStartPosTagOpen>
  <FEndPosOpen>341</FEndPosOpen>
  <FStartPosTagClose>567</FStartPosTagClose>
  <FEndPosTagClose>571</FEndPosTagClose>
  <FParent>211</FParent>
  <FListChild>
       ......
       ......
       ......
  </FListChild>
</node>
```

Figure 7. Node in XML

TABLE I. THE RESULT OF EXTRACTING WEBSITES

Website	Page	Template	Error
http://www.orari.or.id	29	4	0
karnyoto.blogspot.com	54	8	0
www.metrodata.co.id	65	5	0
dprd-pareparekota.go.id	89	3	1
dprd-mamujuutarakab.go.id	280	7	1
www.akprind.ac.id	145	4	3
www.sekolahbogorraya.com	151	8	3
www.sman1-pacitan.sch.id	242	6	1
www.ukitoraja.ac.id	34	4	0
www.ekaristi.org	102	8	2
www.pa-amurang.go.id	221	5	1
www.pa-bitung.go.id	231	4	0
www.dewanpers.or.id	123	6	0
www.sementonasa.co.id	18	3	0
www.stis.ac.id	212	9	0

Fig 7 shows converted node to XML mode, the HTML website tag symbol open (<) and close (>) converting to (~)

symbol because XML parser not read tag symbol of XML and HTML differently.

Table 1 shown that there is still error when doing mapping and extraction on several website pages caused by website programmer’s mistakes in writing tags even though the browser can still read or ignore it.

IV. CONCLUSION

Contents of the BODY element is a combination of news and web tags, a method to extract and mapping web tag into unique node is required, this research using tree node method.

All extracted website has patterns is less than ten patterns that the number of whole page in a website. The least amount of page is 18 and the highest is 280 pages. The system produces less than 2.7% error by using tree node method. The system results which are web tags pattern can already be used as material to determine position of headlines, advertisements and links list, all component position has been listed in unique node. Sometimes, the web programmer is not paying attention to the tags writing in a website. Therefore, there is a tag which has opening but it does not have closing or otherwise. Thus, it is read as “error” by the system. It suggested to having farther research which can obtain the headline of a website by using tree node method and conducting research to format a website tag by using other methods.

ACKNOWLEDGEMENT

This research was fully funded and supported by Toraja Indonesia Christian University.

REFERENCES

[1] 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.

[2] R. Henshaw, " The First Monday Metadata Project," Libri, 1999, pp. 125-131.

[3] S.A. Golder and B.A. Huberman, “Usage Patterns of Collaborative Tagging Systems," Journal of Information Science, vol. 32(2), pp. 198–208, 2006.

[4] T. C. Craven, et al., “HTML Tags as Extraction Cues for Web Page Description Construction," Informing Science Journal, Vol. 6, pp. 1-12, 2003.

[5] 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.

ISBN: 978-602-8509-19-0