A Study on Effective Internet Data Extraction through Layout Detection

Bok-Keun Sun*, Kwang Rok Han
Dept. Computer Engineering, Asan, Hoseo University

ABSTRACT

Currently most Internet documents including data are made based on predefined templates, but templates are usually formed only for main data and are not helpful for information retrieval against indexes, advertisements, header data etc. Templates in such forms are not appropriate when Internet documents are used as data for information retrieval. In order to process Internet documents in various areas of information retrieval, it is necessary to detect additional information such as advertisements and page indexes.

Thus this study proposes a method of detecting the layout of Web pages by identifying the characteristics and structure of block tags that affect the layout of Web pages and calculating distances between Web pages. This method is purposed to reduce the cost of Web document automatic processing and improve processing efficiency by providing information about the structure of Web pages using templates through applying the method to information retrieval such as data extraction.

***Keywords**: Information Retrieval, Data Extraction, Layout, HTML, XML Technologies*

1. INTRODUCTION

Because most Web contents are user-friendly, users who use information regard the user-friendliness as natural. With the explosive growth of Web contents and data, however, documents must be converted into forms that can be processed automatically for structural data process [8]. Accordingly, academic and industrial areas have been studying on methods of abstracting Web data to use in information retrieval. HTML has not been fully standardized and data are user-friendly not machine-friendly, however, there is established tool, technology or standard [7].

Researchers are actively studying methods of processing Web contents mechanically through XML and expect even more active research in the future. Although a relatively late technology compared to HTML, XML is evolving into a language fit for data exchange between heterogeneous machines and automatic data process [14]. However, the number of XML documents accessible on the Web is still insignificant compared to that of HTML documents. Although XML technology will become common on the Web in the future, HTML will be used continuously until XML technology reaches the level of design convenience and user-friendliness of HTML. Thus there are various researches on methods of abstracting data by analyzing HTML documents or converting HTML documents into XML documents, and these researches are considered useful in the areas of information retrieval and data processing. Using these methods, data on websites can be processed automatically by users and applications.

HTML-based Web data are usually developed focused on design and layout rather than on data processing. This makes it difficult to use Web data in data-based information retrieval or calculation or as source data for applications.

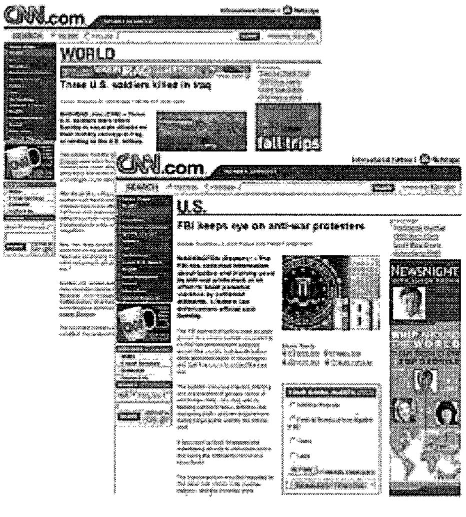

CNN.com
WORLD
Three U.S. soldiers killed in Iraq
CNN.com
U.S.
FBI keeps eye on anti-war protesters
NEWSNIGHT

Fig. 1. Example page using a template.

*Corresponding author. E-mail: bksun@office.hoseo.ac.kr
Manuscript received Sep. 27, 2005 ; accepted Oct. 5, 2005

"This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Governme"

Most sites including news sites that contain data to be used in information retrieval use data presentation structure called template in arranging information to enhance legibility [1][3][9]. As shown in [Figure 1], the main content is positioned at the center and headers, footnotes, advertisements, etc. are placed around. In the template, information is divided visually and semantically.

Pages using the same template have similar layouts and the roles and positions of semantic groups are also similar. Thus, it may be useful for automatic data processing to identify pages of the same layout.

To develop a method of identification, this study examines how to abstract the structure of documents from Web pages. The research is focused on the structure and characteristics of specific HTML tags that affect the layout of Web pages. It calculates distances between Web pages and, if the distance is over a specific threshold, the pages are judged to use the same layout. The developed method is supposed to be applied to information retrieval such as automatic data extraction and provide information about the structure of Web pages using a template, and ultimately, to reduce the cost of Web document automatic processing and enhance processing efficiency.

2. RELATED WORKS

Recently there are increasing interests in researches to analyze the internal structure of Web pages and abstract data. Such researches usually purpose to abstract data through analyzing HTML documents and converting their structure into document forms. Artificial intelligence areas such as machine learning are focused on formulating rules that can abstract sharable general rules through analyzing sample pages and using them for abstracting data from other similar pages [7].

WysiWyg Web Wrapper Factory (W4F) is a tool kit to generate Web wrappers [4]. It includes a unique language to formulate rules for recognizing Web sites and for abstracting data from Web pages. In addition, it provides a mechanism of converting abstracted data into a structure for a specific purpose. The language used here, however, is system-specific one not one compatible with XML.

Web Language (WebL)[10][11] of Compaq is a procedural language for Web wrappers. It provides a powerful abstraction language, mixing regular expressions and reflexive pass expressions, but not compatible with XML technology.

ANDES of IBM uses XML core technologies such as XHTML[12], XPATH[13] and XSLT[16] and provides data abstraction framework [5]. Its approach to XHTML document tree processing uses Xpath (XML Path) and, if necessary, can be combined with regular expressions. The mechanism of XSLT style sheet templates is useful in identifying data extraction patterns represented by the elements of XHTML documents and Xpath, and defines style sheets using a simple description method. However, there are many difficulties in automatic conversion of HTML into XHTML using XSLT including the irregularity of HTML, client scripts and dynamic HTML, so many heuristic methods should be employed. If it is possible to convert all HTML into XHTML correctly using XSLT, any XML tool can be applicable in handling the XHTML pages because XHTML is based on XML [5].

Sometimes annotation-based methods are used in data conversion and extraction. Annotation is a kind of meta data and additional information of Web pages [1] [2] [8][9]. Using annotation, it is possible to abstract and convert information. In general, annotation is related to URI, Xpath and Xpointer[15], which indicate the position of annotation in documents. The author classifies Web pages into a number of semantic groups and save information about each semantic group such as its role and position as an annotation in XML form. Using annotation, it is possible to change the structure of Web pages without changing semantic groups or losing information. Adding annotation into contents, it is possible to convert and abstract contents accurately and precisely. The biggest problem of annotation, however, is cost. It is almost impossible to add annotation to a large volume of frequently updated pages like news. However, most of such Web pages usually use templates. Pages using the same template have similar layouts and the paths and roles of their semantic groups are almost identical.

In order to research Web document data extraction and solve problems in it, this study examines a method of abstracting layout information from Web documents. This studyis focused on the structure and characteristics of specific HTML tags that affect Web page layout. It calculates distances between Web pages and if the distance is below a certain threshold it judges that the pages use the same layout. Through this, this study purposes to provide a method that can be applied to all pages of the same layout in information retrieval areas such as data extraction and annotation making.

3. LAYOUT GROUP DETECTION

In general the main content of a web document is positioned at the center and additional information such as indexes and advertisements are around the main content [Figure 1]. Most documents of this type are created and updated using templates. Some of HTML tags are related to template and layout and information about the tags can be used in calculating distances between pages. First of all, let us discuss the characteristics of HTML tags and then how to calculate distances between pages.

3.1. Block level tag

Block level HTML Tag is a key element in the layout structure of Web pages. [Table 1] shows an example of Block Level Tag.

Table 1. An example of block level tag.

Feature	Block level tag
Table	TABLE, THEAD, TBODY, TR, TH, TD
Form	BUTTON, FORM, TEXTAREA
others	HR, OL, UL, LI, DIV, SPAN, P

Such a tag is called a layout tag in this paper. After analyzing HTML and abstracting layout tags, we structuralize each tag as an Xpath expression and refer it as data source for calculating distances between pages. [Figure 2] shows a HTML document and an example of Xpath abstracted from the document.

3.2. Detection algorithm

To determine if HTML documents use the same template and, as a result, have the same layout, we analyze layout tags as presented above. For this, we convert layout tags into Xpath and calculate distances, and by doing so compare the layouts of Web documents.

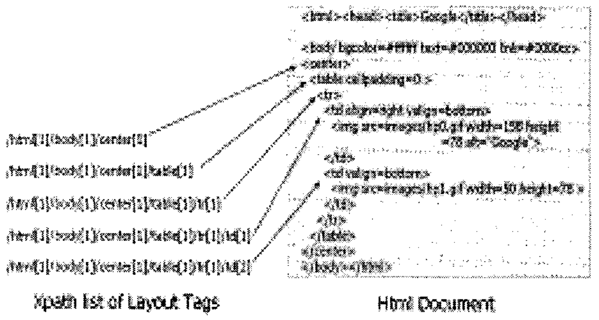


Fig. 2 Xpath expression of a layout tag

When there are two HTML document HDA, HDB, let us say the layout tag lists of the two document HDA, HDB layout as LA and LB. Then the distance (D) between the two documents is calculated using Equation (1).

$$D = \sum_{i} f_i(L_i) \tag{1}$$

where Li , LA, LB and fi is a function to calculate the distance between LA and Li, which is the ith layout of LB. The distance calculating function fi is as Equation (2).

$$f_i(L_i) = \begin{cases} 0 & L_i \in (L_A \cap L_B) \\ W_i * L_i & L_i \notin (L_A \cap L_B) \end{cases} \tag{2}$$

In Equation (2) above, if the layout tags expressed in document HDA and HDB, namely, Xpaths are identical with each other 0 is returned, and if not a positive number is returned. When Xpaths are identical with each other the values of attributes such as bgcolor and alignin the corresponding tags may be compared, but in this study if Xpaths are identical with each other the distance calculating function simply returns 0 [6]. Wi is the weight of a tag, and Li is a parameter value. Using the equation above, this study calculates the distance between Web document HDA and HDB. If the result is not larger than T, a specific threshold, the two documents are judged to have the same layout.

3.3 Parameter setting

Weight parameter Wi was set at 1. The weights of layout tags were set using a heuristic method as shown in Table 2.

Table 2. The values of L parameters.

Layout Tag	L Value
THEAD,TBODY,HR	1
other block tags	infinite

Parameter values were set so that if Xpaths are not identical with each other the structures of the pages are judged to be different from each other except the corresponding tag is THEAD, TBODY or HR.

4. IMPLEMENTATION

This study implemented a HTML analysis too and HALD (HTML Analyzer for Layout Detection) system based on the algorithm and parameter values described in Chapter 3 to evaluate the performance of the algorithm and the system. [Figure 3] and [Figure 4] show the outline of the HTML analysis tool system and a screen executing the system.

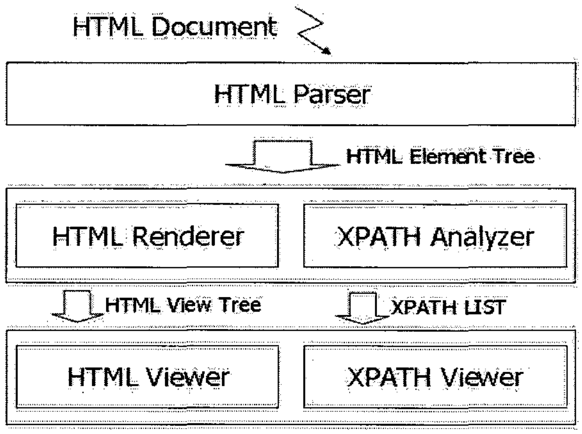


Fig. 3. Outline of HTMLanalysis tool system

Fig. 4. Screen executing HTML analysis tool.

HTML parser in [Figure3] parses documents and creates a HTML element tree similar to DOM tree. Using the element tree, the tree generator creates view trees and Xpath analyzer creates Xpath lists. The left of [Figure 4] shows a screen executing the viewer and the right shows a screen executing Xpath analyzer.

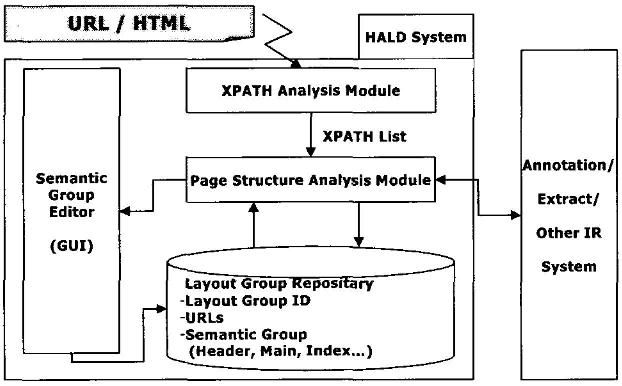


Fig. 5. Outline of HALD system

Xpath analyzer is included in Xpath analysis module within HALD system in [Figure 5]. The structural diagram of HALD system to be implemented for data extraction system is as in [Figure 5].

Xpath analysis module in [Figure 5]abstracts Xpath lists from input HTML documents and page structure analysis module uses the lists as input and calculates distance to Xpath in storage. If the module judges that a page of the same structure is in the storage it assigns the corresponding semantic group to the HTML document and save it. If the module fails to find a page of the same structure, it assigns a new semantic group through semantic group editor and registers the page in the storage as a document of a new structure.

Systems in information retrieval areas such as data extraction can perform information retrieval effectively using HALD through API provided by the page structure analysis module.

5. PERFORMANCE EVALUATION

This chapter analyzes the structure of Web pages using the algorithm and equations presented in Chapter 3 and evaluate the results. First, [Table 3] is the result of a test using 5 and 10 as threshold T. According to the result, World News of CNN , Latest News of Yahoo News and Politics News of Kyunghyang Shinmun were all created using templates and each site used the same template to create its documents.

Table 3. Result of document structure analysis

Web Site	T = 5		T = 10	
www.cnn.com	classifying	recall	classifying	recall
news.yahoo.com	13	32.5	15	37.5
www.khan.co.kr	5	12.5	8	20
Total	30	75	30	75

Forty articles in each site all used the same template, and the test was carried out by measuring recall values. According to the result of the test, when documents contain relevant article boxes inserted using <DIV>, <TABLE>, etc. the system often judged that their structures are different.

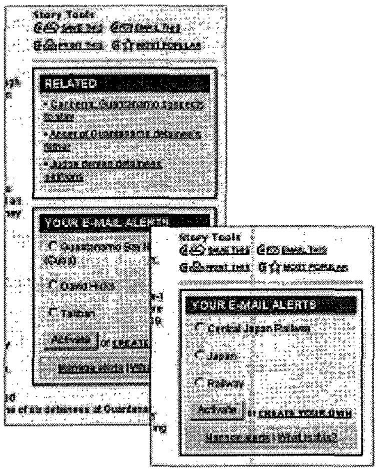


Fig. 6. Examples of relevant article boxes in CNN News site.

Fig. 6 shows examples of pages that used the same template but were judged to have different structure as they had relevant article boxes. When a heuristic method of excluding relevant article boxes tagged with <DIV>, <TABLE>, etc. was applied and structure analysis was carried out again on the pages, the result was as shown in [Table 4], which shows that the application of the heuristic method improved the result to some degree.

Table 4. Result of analysis applying a heuristic method.

Web Site	T = 5		T = 10	
www.cnn.com	classifying	recall	classifying	recall
news.yahoo.com	19	47.5	21	52.5
www.khan.co.kr	15	37.5	18	45
www.khan.co.kr	38	95	38	95

6. CONCLUSIONS AND FUTURE WORKS

Web services and designs are growing more complicated for instance, a system showing users' opinion on articles and they are making automatic data processing more difficult. The development of many heuristic methods, however, is expected to resolve such difficulties.

Many Web pages including news are created using data presentation structures called templates. Thus this study developed and tested a system that analyzes the structure of web pages and applies the results in various areas including information retrieval. This study developed first HTML analysis tool and Xpath analyzer and tested their performance. This study will be continued to develop Xpath analyzer and HALD system of upgraded performance through improving the algorithm, applying various parameter values and discovering additional heuristic methods and apply them to information extraction and processing in information retrieval areas.

REFERENCE

[1] C. Asakawa, H. Takagi, "Annotation-Based Transcoding for Nonvisual Web Access" Proceedings of ACM ASSETS 2000, pp.172-179, Nov 2000.

[2] M. Hori, G. Kondoh, K. Ono, S. Hirose and S. Singhal, "Annotation-Based Web Content transcoding", Proceedings of the 9th International WWW Conference, pp.197-211, May 2000.

[3] T. Sullivan, R.Matson, "Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites" Proceedings of ACM CUU 2000, pp.139-144,Nov 2000.

[4] A. Sahuguet and F. Azavant, "Building Light-Weight Wrappers for Legacy Web Data-Sources Using W4F", International Conference on Very Large Data Bases(VLDB), Edinburgh, Scotland, Sep 1999.

[5] Myllymaki J, "Effective Web data extraction with standard XML technologies", The International Journal of Computer & Telecommunications, V.39 N.5, 635-644, AUG, 2002.

[6] K. Fukuda, H. Takagi, J. Maeda and C. Asakawa "Layout Group Extraction from Web Content for Effective Adaptation", IBM Research Report, 14 Nov 2002.

[7] J. Myllymaki and J. Jackson, "Robust Web Data Extraction with XML Path Expressions", IBM Research Report, 23 May 2002.

[8] K. Nagao, Y. Shirai and K. Squire, "Semantic Annotation and Transcoding : Making Web Content More Accessible" IEEE MultiMedia, vol. 8, pp.69-81, Apr 2001.

[9] A. W. Huang and N. Sundaresan, "A Semantic Transcoding System to Adapt Web Services for Users with Disabilities" Proceedings of ACM ASSETS 2000, pp.156-163, Nov 2000.

[10] C. Allen, "WIDL:Application Integration with XML", WWW Journal2(4), Nov 1997.

[11] Web Interface Definition Language, W3C Note. Sep 1997, "http://www.w3c.org/TR/NOTE-WIDL.

[12] XHTML, W3C Recommendation, May 2003. Http://www.w3c.org/TR/xhtml2.

[13] XML Path Language (Xpath), W3C Recommendation, Nov 1999. Http://www.w3c.org/TR/xpath.

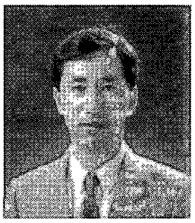
[14] Extensible Markup Language (XML), W3C Recommendation, Feb 1998. Http://www.w3c.org/XML.

[15] XML Pointer (Xpointer) W3C Recommendation, http://www.w3c.org/XML/Linking, Sep 2001.

[16] XSL Transformations (XSLT) W3C Recommendation, http://www.w3c.org/TR/xslt, Nov 1999.

BokKeun Sun

He received the B.S, M.S degrees in computer engineering from hoseo University, Cheonan, Korea in 1999 and 2001. And He has been taking PH.D course in Computer Engineering in Hoseo University since 2003. He is specially intereted in Information Retrieval, Simulation, and Agent.



KwangRok Han

He received the B.S, M.S, and Ph.D degrees in electronics and information engineering from Inha university, Inchon, Korea in 1984, 1986 and 1989.

He was a senior researcher at Korea Institute of Sports Science until 1991 from 1989. He is a professor of dept of Computer Engineering in Hoseo University from 1991 and a visiting professor of ISI Uiniversity of South California University.