Partitioning Based Web Content Mining

Niki R. Kapadia
nikikapadia4@gmail.com
ME CSE, GEC Modasa

Kanu Patel
kanu.patel@live.com
ME CSE, GEC Modasa

Mehul C.Parikh
Prof.mehulcparikh@gmail.com
Assistant Professor,GEC-Modasa

***Abstract*—** *Today the Web has become the largest information source for people. Most information retrieval systems on the Web consider web pages as the smallest and undividable units, but a web page as a whole may not be appropriate to represent a single semantic. A web content structure analysis based on visual representation is proposed in this dissertation work. Many web applications such as information retrieval, information extraction and automatic page adaptation can benefit from this structure. Furthermore, web page often contains multiple topics that are not necessarily relevant to each other. Therefore, detecting the semantic content structure of a web page could potentially improve the performance of web information retrieval. This dissertation work presents an automatic top-down, tag-tree independent approach to detect web content structure. It simulates how a user understands web layout structure based on his visual perception and information can be mined dynamically. Comparing to other existing techniques, our approach is independent to underlying documentation representation such as HTML.*

***Keywords*— Data mining, web mining, web content mining, web structure mining, web usage mining, dustering, segmentation.**

I. INTRODUCTION

Web mining is the data mining technique that automatically discovers/extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

mining techniques in detail, results and comparison to extract necessary information effectively and efficiently.

II. WEB MINING PROCESS

Web mining process is shown in figure 2. And its steps are given below:

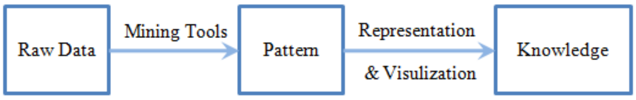


Figure 1 web mining process

Steps of web mining process:

a. Resource Finding: - It is the task of retrieving intended web documents.
b. Information selection and preprocessing:- Automatically selecting and pre-processing specific from information retrieved web resources.
c. Generalization:-Automatically discovers general patterns at individual web sites or multiple sites.
d. Analysis:-Validation and interpretation of the mined patterns.

III. WEB MINING CATEGORIES

Web mining can be categorized as below.

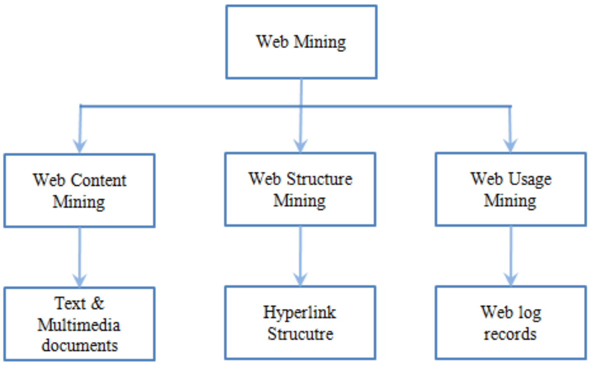


Fig 2. Web mining categories

a. **Web Content Mining:** - Web content mining is the process of extracting useful information from the contents of web documents. It is related to data mining. It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. Web content mining is semi-structured nature of the web. Technologies used in web content mining are NLP, IR.
b. **Web Structure Mining:** - tries to discover useful knowledge from the structure and hyperlinks. The goal of web structure mining is to generate structured summery about websites and web pages. It is using tree-like structure to analyze and describe HTML or XML.
c. **Web Usage Mining:** - Web usage mining is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on technique that can be used to predict the user behavior while user interacts with the web. It uses the secondary data on the web. This activity involves automatic discovery of user access patterns from one or more web-servers. It consists of three phases namely: pre-processing, pattern discovery, pattern analysis. Web servers, proxies and client applications can quite easily capture data about web usage.

IV. METHOD TO MINE DATA

The vision-based content structure of a page is obtained by combining the DOM structure and the visual cues. It is the combination of hierarchical and partitioning based method. The segmentation process is illustrated in figure 3. It has mainly three steps: (a) block extraction, (b) separator detection and (c) content structure construction. These three steps as a whole are regarded as a round. The algorithm is top-down. The web page is firstly segmented into several big blocks and the hierarchical structure of this level is recorded. For each big block, the same segmentation process is carried out recursively until we get sufficiently small blocks.

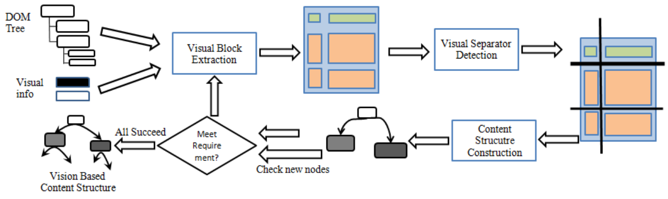

DOM Tree
Visual info
Visual Block Extraction
Visual Separator Detection
Content Strucutre Construction
Check new nodes
Meet Require ment?
All Succeed
Vision Based Content Structure

Figure 3 The vision-based page segmentation algorithm

a. Visual Block Extraction.

In this step, we aim at finding all appropriate visual blocks contained in the current subpage [16]. In general, every node in the DOM tree can represent a visual block. However, some "huge" nodes such as <TABLE> and <P> are used only for organization purpose and are not appropriate to represent a single visual block. In these cases, the current node should be further divided and replaced by its children. Due to the flexibility of HTML grammar, many web pages do not fully obey the W3C HTML specification, so the DOM tree cannot always reflect the true relationship of the different DOM node.

For each extracted node that represents a visual block, its DoC value is set according to its intra visual difference. This process is iterated until all appropriate nodes are found to represent the visual blocks in the current sub-page.

```
Algorithm DivideDomtree(pNode,nLevel)
 {
IF(Dividable(pNode,nLevel)==TRUE)
   FOR EACH child OF pNode {
     DivideDomtree(child,nLevel);
   }
} ELSE {
Put the sub-tree (pNode) into the pool as a block;
   }
}
```

Figure 4 The visual block extraction algorithm

b. Visual Separator Detection.

After all blocks are extracted, they are put into a pool for visual separator detection. Separators are horizontal or vertical lines in a web page that visually cross with no blocks in the pool. From a visual perspective, separators are good indicators for discriminating different semantics within the page. A visual separator is represented by a start pixel and end pixel. The width of the separator is calculated by the difference between these two values.

c. Separator Detection

The visual separator detection algorithm is described as follows:

Firstly the separator list will be initialized. The list starts with only one separator, whose start pixel and end pixel are corresponding to the borders of the pool. For every block in the pool, the relation of the block with each separator will be evaluated. Then remove the separators that stand at the border of the pool.

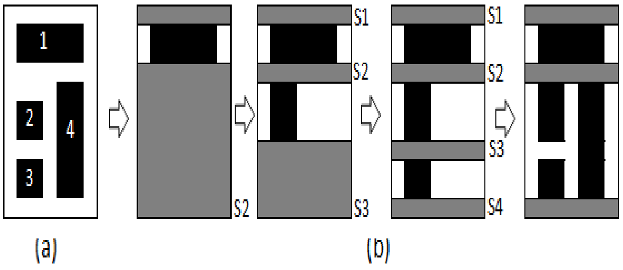


Figure 5 A sample page and the separator detection process [16]

Take figure 5(a) as an example in which the black blocks represent the visual blocks in the page. For simplicity we only show the process to detect the horizontal separators. At first we have only one separator that is the whole pool. As shown in figure 5(b), when we put the first block into the pool, it splits the separator into S1 and S2. It is the same with the second and third block. When the fourth block is put into the pool, it crosses the separator S2 and covers the separator S3, the parameter of S2 is updated and S3 is removed. At the end of this process, the two separators S1 and S3 that stand at the border of the pool are removed.

d. Content Structure Construction

When separators are detected and separators' weights are set, the content structure can be constructed accordingly [16]. The construction process starts from the separators with the lowest weight and the blocks beside these separators are merged to form new blocks. This merging process iterates till separators with maximum weights are met.

After that, each leaf node is checked whether it meets the granularity requirement. For every node that fails, we go to the Visual Block Extraction step again to further construct the sub content structure within that node. If all the nodes meet the requirement, the iterative process is then stopped and the vision-based content structure for the whole page is obtained. In summary, the proposed VIPS algorithm takes advantage of visual cues to obtain the vision-based content structure of a web page and thus successfully bridges the gap between the DOM structure and the semantic structure. The page is partitioned based on visual separators and structured as a hierarchy. This semantic hierarchy is consistent with human perception to some extent. VIPS is also very efficient. Since we trace down the DOM structure for visual block extraction and do not analyze every basic DOM node, the algorithm is totally top-down.

V. FLOW OF IMPLEMENTATION

We will illustrate example of vision-based content structure for web page actually used for project purpose, IDOM technique, dataset, coding language, database software etc. We are using web source "http://twibs.com/alphabetical.php" for mining the web data. The entire process from extracting of web data to content structure generation is elaborately discussed.

.

a. Web Server and Database

WAMP, used for project purpose is a Windows OS based program that installs and configures Apache web server, MySQL database server, PHP scripting language, phpMyAdmin (to manage MySQL database's), and SQLiteManager (to manage SQLite database's). WAMP is designed to offer an easy way to install Apache, PHP and MySQL package with an easy to use installation program instead of having to install and configure everything yourself.

MySQL is currently the most popular open source database server in existence. On top of that, it is very commonly used in conjunction with PHP scripts to create powerful and dynamic server-side applications. MySQL has been criticized in the past for not supporting all the features of other popular and more expensive Database Management Systems and it has become widely popular with individuals and businesses of many different sizes.

b. Programming Language-PHP

PHP is a general purpose server-side scripting language originally designed for web development to produce Dynamic Web page[15]. It is one of the first developed server-side scripting languages to be embedded into an HTML source document, rather than calling an external file to process data. Ultimately, the code is Interpreter by a Web server with a PHP processor module which generates the resulting Web page. PHP can be deployed on most Web servers and also as a standalone Shell on almost every Operating system and Platform free of charge. A competitor to Microsoft's Active Server Pages (ASP) server-side script engine and similar languages, PHP is installed on more than 20 million Web sites and 1 million Web servers.

PHP is an HTML-embedded scripting language. Much of its syntax is borrowed from C, Java and Perl with a couple of unique PHP-specific features thrown in. The goal of the language is to allow web developers to write dynamically generated pages quickly. When someone visits your PHP webpage, your web server processes the PHP code. It then sees which parts it needs to show to visitors (content and pictures) and hides the other stuff (file operations, math calculations, etc.) then translates your PHP into HTML. After the translation into HTML, it sends the webpage to your visitor's web browser.

PHP will allow you to:

- Reduce the time to create large websites.
- Create a customized user experience for visitors based on information that you have gathered from them.
- Open up thousands of possibilities for online tools. Check out PHP - HotScripts for examples of the great things that are possible with PHP.
- Allow creation of shopping carts for e-commerce websites.

c. Document Object Model(DOM)

The Document Object Model is a platform and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. The document can be further processed and the results of that processing can be incorporated back into the presented page. "Dynamic HTML" is a term used by some vendors to describe the combination of HTML, style sheets and scripts that allows documents to be animated.

d. VIPS Algorithm

Input: A set of web pages (W) from a given website http://twibs.com/alphabetical.php (screen shot as shown in figure 5.2), maximum number of blocks for outputting in each web page (N).

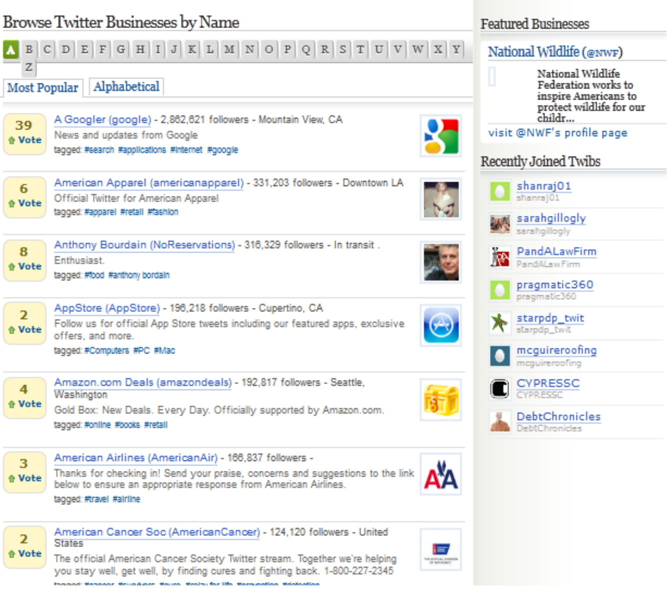
Output: A data set containing plain text as well as in

Tabulated form from web pages.

Begin

1. Apply vision based partitioning algorithm using IDOM technique to segment web pages, (W) into blocks .
2. Data Extraction and storing of data set into database.
3. Generate cleaned files records.

End



Browse Twitter Businesses by Name

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Most Popular | Alphabetical

39 Vote — A Googler (google) - 2,862,621 followers - Mountain View, CA
News and updates from Google
tagged: #search #applications #internet #google

6 Vote — American Apparel (americanapparel) - 331,203 followers - Downtown LA
Official Twitter for American Apparel
tagged: #apparel #retail #fashion

8 Vote — Anthony Bourdain (NoReservations) - 316,329 followers - In transit .
Enthusiast.
tagged: #food #anthony bordain

2 Vote — AppStore (AppStore) - 196,218 followers - Cupertino, CA
Follow us for official App Store tweets including our featured apps, exclusive offers, and more.
tagged: #Computers #PC #Mac

4 Vote — Amazon.com Deals (amazondeals) - 192,817 followers - Seattle, Washington
Gold Box: New Deals. Every Day. Officially supported by Amazon.com.
tagged: #online #books #retail

3 Vote — American Airlines (AmericanAir) - 166,837 followers -
Thanks for checking in! Send your praise, concerns and suggestions to the link below to ensure an appropriate response from American Airlines.
tagged: #travel #airline

2 Vote — American Cancer Soc (AmericanCancer) - 124,120 followers - United States
The official American Cancer Society Twitter stream. Together we're helping you stay well, get well, by finding cures and fighting back. 1-800-227-2345

Featured Businesses

National Wildlife (@NWF)

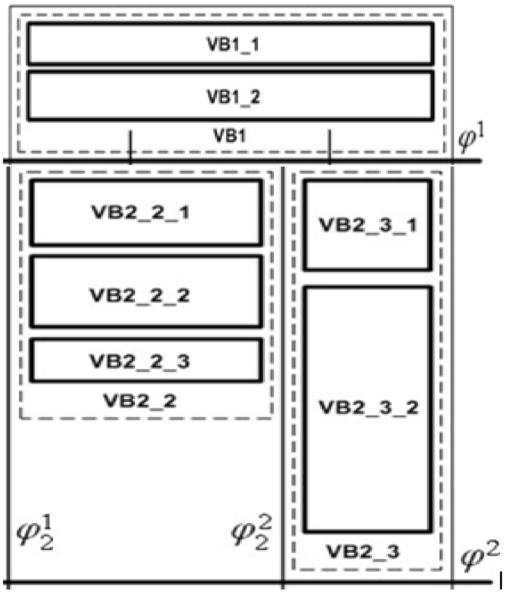
National Wildlife Federation works to inspire Americans to protect wildlife for our childr...

visit @NWF's profile page

Recently Joined Twibs

shanraj01 shanraj01
sarahgillogly sarahgillogly
PandALawFirm PandALawFirm
pragmatic360 pragmatic360
starpdp_twit starpdp_twit
mcguireroofing mcguireroofing
CYPRESSC CYPRESSC
DebtChronicles DebtChronicles

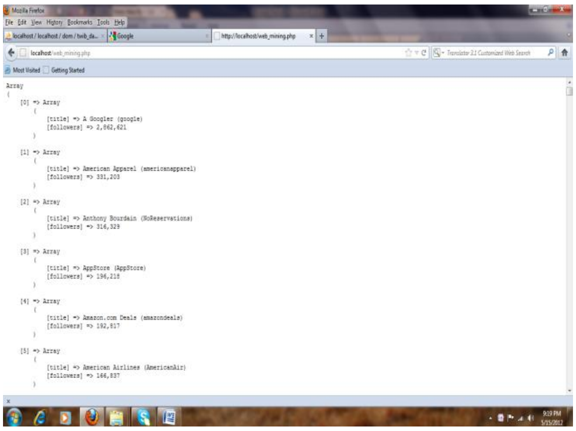
(a)



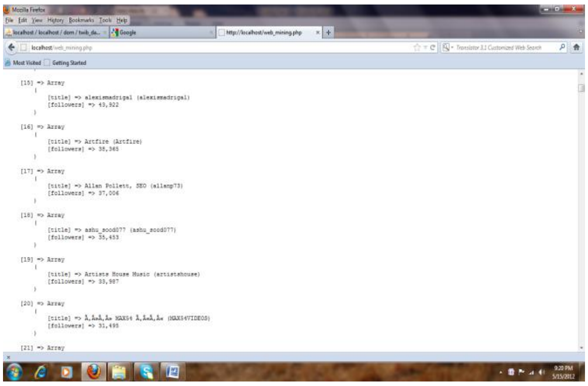
(b)

Figure 6 (a) Website Screenshot (b) Vision-based Content Structure.

e. Resultant Data set.


Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost / localhost / dom / twib_da...
Google
http://localhost/web_mining.php
localhost/web_mining.php
Most Visited
Getting Started

```
Array
(
    [0] => Array
        (
            [title] => A Googler (google)
            [followers] => 2,862,621
        )

    [1] => Array
        (
            [title] => American Apparel (americanapparel)
            [followers] => 331,203
        )

    [2] => Array
        (
            [title] => Anthony Bourdain (NoReservations)
            [followers] => 316,329
        )

    [3] => Array
        (
            [title] => AppStore (AppStore)
            [followers] => 196,218
        )

    [4] => Array
        (
            [title] => Amazon.com Deals (amazondeals)
            [followers] => 192,817
        )

    [5] => Array
        (
            [title] => American Airlines (AmericanAir)
            [followers] => 166,837
        )
```


Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost / localhost / dom / twib_da...
Google
http://localhost/web_mining.php
localhost/web_mining.php
Most Visited
Getting Started

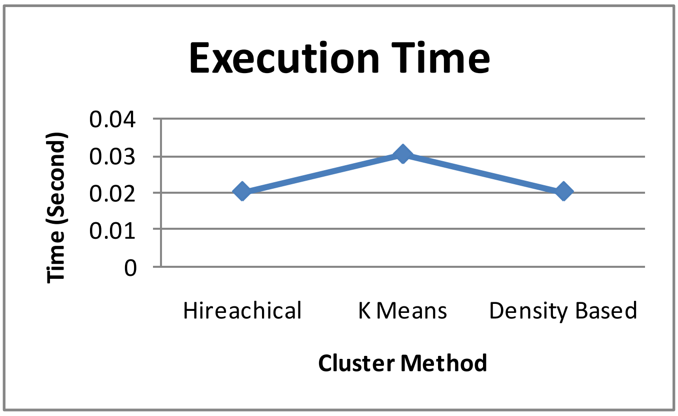
```
    [15] => Array
        (
            [title] => alexismadrigal (alexismadrigal)
            [followers] => 43,922
        )

    [16] => Array
        (
            [title] => Artfire (Artfire)
            [followers] => 38,365
        )

    [17] => Array
        (
            [title] => Allan Pollett, SEO (allanp73)
            [followers] => 37,006
        )

    [18] => Array
        (
            [title] => ashu_sood077 (ashu_sood077)
            [followers] => 35,453
        )

    [19] => Array
        (
            [title] => Artists House Music (artistshouse)
            [followers] => 33,987
        )

    [20] => Array
        (
            [title] => Å‚ÅœÅ‚Å» MAX54 Å‚ÅœÅ‚Å« (MAX54VIDEOS)
            [followers] => 31,495
        )

    [21] => Array
```

VI. RESULT ANALYSIS.

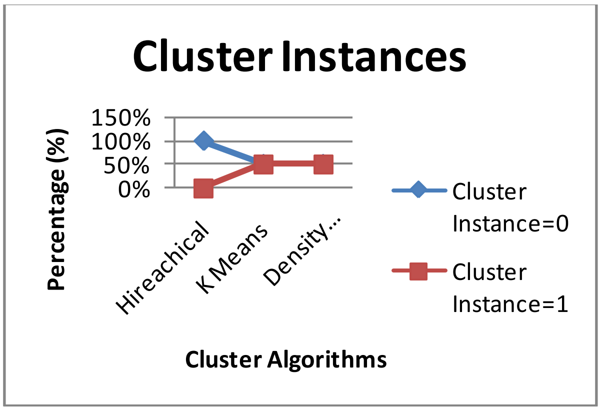
For the analysis of the results, WEKA tool has been used.

The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the weka software. Each of the major weka packages Filters, Classifiers, Clusters, Associations, and Attribute Selection is represented in the Explorer along with a Visualization tool which allows datasets and the predictions of Classifiers and Clusters to be visualized in two dimensions.

In WEKA tool, analysis of different clustering algorithms such like Hierarchical, K Means and density based are performed. The time taken by these algorithms are mapped and shown in graphs. and shown in the below graphs.



Graph-1 Execution Time for various cluster algorithms



Graph-2 Result of Cluster Instance

It is easier to mine the website data using php script language and wamp server. The same has been used for dissertation work. The clean data as desired is mined from the website. The mined data is downloaded in *.csv format from wamp server for analysis. The weka tool is used for the analysis, using different cluster methods. The generated instances and execution time for different cluster methods represent that hierarchical cluster method takes higher execution time and higher 0 instances with respect to K Means and Density based cluster algorithms.

VII. CONCLUSION

An approach for extracting web content structure based on visual representation is proposed. The resulted web content structure is very helpful for applications such as web adaptation, information retrieval and information extraction. By identifying the logic relationship of web content based on visual layout information, web content structure can effectively represent the semantic structure of the web page. An automatic top-down, tag-tree independent and scalable algorithm to detect web content structure is presented. It simulates how a user understands the layout structure of a web page based on its visual representation. Compared with traditional DOM based segmentation method, our scheme utilizes useful visual cues to obtain a better partition of a page at the semantic level. The algorithm is evaluated manually on a large data set, and also used for selecting good expansion terms in a pseudo-relevance feedback process in web information retrieval, both of which achieve very satisfactory performance.

Recently, the developed algorithm is implemented on website with horizontal separation. The further work would be focused on development of improved code in order to take care of vertical separation also. Further, the various experiments of developed code will be carried out on different domain website such like education, shopping, social networking etc. to check accuracy of the code and to judge which consideration will give better results.

This work will be useful in understanding VIPS algorithm, web content mining through partition based segmentation and further research directions in this area to computer engineering fraternity.

VI. REFRENCES

[1] Free encyclopedia. (2012, May 13). Data Mining[Online]. Available: http://en.wikipedia.org/wiki

[2] Bing Liu,"Web Data Mining",1st Edition,Springer,July 2011.

[3] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques",2nd edition, Morgan Kaufmann Publishers, March 2006.

[4] Alex Berson, Stephen Smith, and Kurt Thearling. (2010). An Overview of Data Mining Techniques [Online]. Available: http://www.thearling.com

[5] Web Mining Techniques, Tools and Data Ware housing. Available: http://www.web-datamining.net

[6] Soumen Chakrabart,"Mining the Web", Morgan Kaufmann publisher, Part-II.

[7] Margaret H.Dunham,"Data Mining",Pearson Education,2nd Edition,2007.

[8] http://www.loginworks.com/web-data-mining

[9] Markus Schedl1, Peter Knees1, Tim Pohle1, and Gerhard Widmer, "Towards an Automatically Generated Music Information System via Web Content Mining",in Proceedings International journal of Information Processing & Management 47 in 2011, pp.426-439.

[10] Lihui Chen, Wai Lian Chue, (2004), "Using Web structure and summarization techniques for Web Content mining", in Proceedings International journal of Information Processing & Management 41 in 2005, pp..1225-1242.

[11] Bing Liu, Kevin Chen-chuan Chang, "Editorial: Special issue on Web content Mining", SIGKDD Explorations, Volume 6, Issue 2.

[12] Jing Li and C.I. Ezeife, "Cleaning Web Pages for Effective Web Content mining", in proceedings: DEXA, Published in 2006.

[13] Dorian Pyle,"Data Preparation for data mining", Morgan Kaufmann Publishers, March 1999.

[14] Ji-Rong Wen, Wei-Ying Ma. (2003). VIPS: a Vision-based Page Segmentation. Algorithm [Online]. Available FTP: ftp://ftp.research.microsoft.com/pub/tr/tr-2003-79.pdf.

[15] The PHP Group. (2001). What is PHP? [Online]. Available: http://www.php.net

[16] Deng ci, Sjipeng Yu, et. al., "VIPS: A Vision-based Page Segmentation Algorithm", Redmond, WA 98052, Nov. 1, 2003.