Extraction of Informative Blocks from Myanmar Web Pages Based on Entropy Measure

Swe Swe Nyein
University of Computer Studies, Mandalay
sweswenyein@gmail.com

Abstract

With the large amount of information on the Internet, Web pages have been the potential source of information retrieval and data mining technology. Apart from the informative blocks (main blocks), a web page also comprises of noisy information that can degrade the performance of information retrieval applications. A method to identify and extract the informative content from web pages is needed. In this paper, we propose an Informative Block Extraction System using Document Object Model (DOM) tree and Entropy Evaluation Model (EEM) which quantifies the expected value of the information contained in a web page. To evaluate the proposed system, Myanmar News Web Pages are applied and also measure the accuracy of the system, precision and recall are used.

1. Introduction

The World Wide Web (WWW) rapidly grows as it is accessible for public use through the web browser. Typically, Myanmar Web pages comprise of different kinds of contents. i.e. a News page besides the article posting as the main content it also contains other noisy contents such as navigation bar, directory lists, advertisements, user comments, header and footer. When browsing a web page, most of the time users typically focus on the main content and ignore the additional contents (noisy contents). For human, this behavior can be done relatively fast and accurate because they can use their knowledge, visual representation and layout of the web pages to distinguish the main content from other parts.

In the other hand, since computer software is not as intelligent as human to distinguish between the main content and the noisy content, this become the challenge for commercial search engines, web miners and other kinds of applications that use web documents as a data source. A search engine typically indexes the whole text of a web page. As a result, the noisy data which is useless information remains in the index and also these data may degrade the accuracy of the search results as well as search speeds, information extraction and the size of the index. For rapid and convenient access to the useful data contained in the Web page, an extraction technique is required.

To extract the informative contents, the system needs to detect the noisy contents correctly. While noise filtering algorithms are usually used to improve the accuracy of induce classification models, the system aim to distinguish the main content from unrelated information that are negatively affects users of small display devices such as PDAs (Personal Digital Assistants) and cell phone but they are functionally useful human browsers and necessary for web site owners.

This paper proposes a system which consists of a four-step strategy to extract the informative contents: Preprocessing, DOM tree, Entropy Measure and Block Identification. A parser is needed to parse web documents (i.e DOM parser). The system doesn't use the entire DOM tree to evaluate entropy because a DOM tree is directly built by html code, the tree structure is extremely complex. So it is needed to filter unnecessary tags and also segment the DOM tree.

This paper is organized in the following sequence. A briefly survey of many researchers is presented in section 2. The proposed system is dealt with in section 3. This is followed by the evaluation and conclusion in section 4 and 5.

2. Related Works

Detecting and extracting main block from Web pages is an important problem. Various techniques have been developed to deal with this problem.

T. Win and K. N. N. Tun [9] proposed rules based approach to eliminate the noise and defined the rules which are noise in web page.

T. Htwe [8] presented an approach NoiseEliminator that detected multiple noise patterns and removed these from web pages based on Case-Based Reasoning (CBR) and back propagation neural network algorithm.

L. Yi et al. have proposed a new Style tree to capture the actual contents and common layouts (or presentation styles) of the Web pages in a Web site. Their method can difficult to capture the common presentation style for many web pages from different web sites in [3].

Another approach mentioned in S. H. Lin and J. M .Ho [7] was InfoDiscoverer system to discover informative content blocks from web documents. It first partitions a web page into several content blocks according to HTML tag <TABLE>. They considered only <TABLE> tag for blocking.

The approach described in C. Li et al. [1] extracted informative block from a web page based on the analysis of both layouts and semantic information of the web pages. They needed to identify blocks occurring in a web collection based on the Vision-based Page Segmentation algorithm.

In [5], P. S. Hiremath et al. proposed an algorithm called VSAP (Visual Structure based Analysis of web Pages) to exact the data region based on the visual clue (location of data region / data records / data items / on the screen at which tag are rendered) information of web pages. In [2] D. Cai et al. proposed a Vision-based Page Segmentation (VIPS) algorithm that segments web pages using DOM tree with a combination of human visual cues, including tag cue, color cue, size cue, and others.

P. M. Joshi et al. proposed an approach of combination of HTML DOM analysis and Natural Language Processing (NLP) techniques for automated extractions of main article with associated images form web pages. Their approach did not require prior knowledge of website templates and also extracted not only the text but also associated images based on semantic similarity of image captions to the main text in [4].

In [11], Y. Li and J. Yang proposed a tree called content structure tree which captured the importance of the blocks. In [6], S. Gupta et al. proposed content extraction technique that could remove clutter without destroying webpage layout. It is not only extract information from large logical units but also manipulate smaller units such as specific links within the structure of the DOM tree. In [10], Y. Fu et al. proposed a technique to discover informative content block based on DOM tree. They removed clutters using XPath. They removed only the web pages with similar layout.

Although most of the existing approaches have been solved content extraction problem, they did not solve Informative Block Extraction problem from Myanmar Web pages. In this paper, we intend to extract the informative blocks from Myanmar Language Web pages.

3. System Architecture

The proposed architecture includes four stages to handle the Web pages: Preprocessing, DOM tree, Entropy Evaluation Model and Block Identification. The system transforms Web pages into DOM tree, Entropy Evaluation Model evaluates entropy of nodes according to the trees and block type is identified according to the entropy value such as informative block or non-informative block.

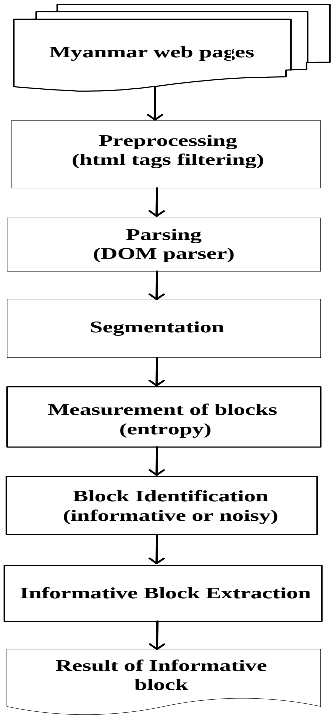
Myanmar web pages
Preprocessing (html tags filtering)
Parsing (DOM parser)
Segmentation
Measurement of blocks (entropy)
Block Identification (informative or noisy)
Informative Block Extraction
Result of Informative block

Figure 1. Overview of the Proposed System

3.1. Preprocessing

The system filters tags that are unable to appear on web browser such as comment, script, and style and so on. For example;
<div class = “clear”> </div>
<div style = “ display: none;”> </div>
<style> <!------- menu jq -------> </style>

Most of the Web pages are not well-formed. To construct the DOM tree, we need to check web documents using html tidy tool. In html code, a tag contained by < and >. Each of them has a name and a scope indicating a region from start tag and end tag. Some tags such as <img>, <input>, <br> and <hr> are not required end tag in html document but the start tag and end tag are needed to build the DOM tree. So, tidy tool can detect, add and correct missing or mismatched end tags and so on. The valid html document is transformed into DOM tree.

3.2. DOM Parser

A wrapper normally transforms a Web page into a certain kind of data structure for conveniently and easily accessing data. There are various kinds of parsers such as SAX parser, html parser and Pull parser. SAX parser can’t insert or delete a node. In my proposed system, nodes are needed to remove. So the system is convenient with the DOM parser. DOM is a well-known model to present a web page, and a modeled page is viewed as a tree object, DOM tree. DOM parser transforms a web page into a DOM tree, which it exploits tags within a page and builds a tree object on relationship between tags. According to the DOM, everything in an HTML document is a node. HTML Web pages begin from the BODY tag since all the viewable parts are within the scope of BODY. The Figure 2 is transformed into DOM tree as shown in Figure 3.

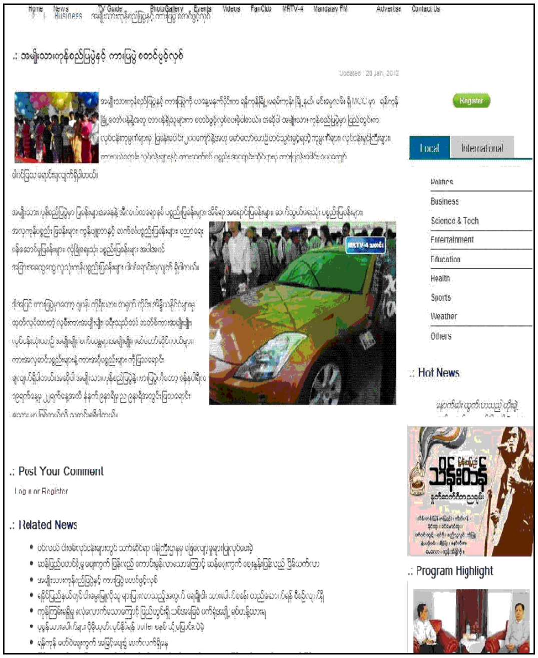

Figure 2. Sample of Myanmar News page

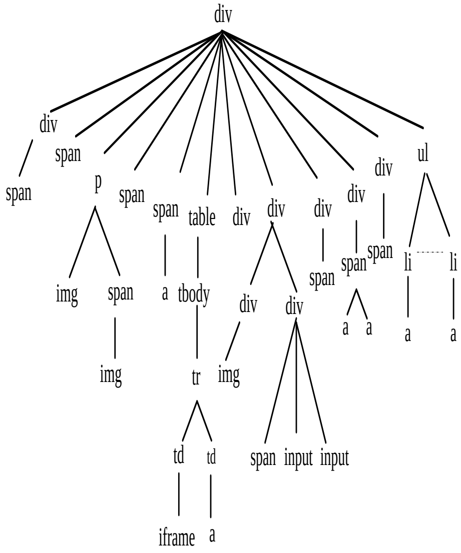

Figure 3. DOM tree of Sample Web page

3.3. Page Segmentation

The system doesn't use the entire DOM tree to evaluate entropy because a DOM tree is directly built by html code, the tree structure is extremely complex. DOM tree is segmented into sub-trees with the specific tags of html. Each sub-tree is a block. We classify each HTML tag element to be either a “block" or “style" element. Each block element, such as the <p>, is rendered into a content block on the web page. It impacts the page layout or the relative positioning of the content. Style elements, in contrast, only affect the visual attributes of the content, such as the font size or color but not the layout. We define block tags to be the following set (all others are considered as style tags): div, p, br, li, ul, ol, td, tr, table, h1-6, hr.

3.4. Entropy Evaluation Model (EEM)

The entropy is very appropriate for measuring the uncertainty and the degree of the information content of the block or sub-tree can indicate the block characteristic. We measure the weight of the element nodes (such as text, image, and link) appearing in a block based on the entropy technique. Weight is the total number of links, images and text in each block. The following eq. (1) is Shannon's famous general formula for uncertainty:

$$0 \le H = -\sum_{i=1}^{n} p_i \log_2 p_i \le \log_2 n , \qquad (1)$$

where p_i is the probability of event.

Based on the eq. (1), we normalize the weight of the node; the node (block) entropy is as shown in following equation:

$$H(block) = -\sum_{i=1}^{n} \frac{(EN)_i}{W} \log_2 \frac{(EN)_i}{W} , \qquad (2)$$

Where,

-n is the number of type of element node (img, link, text).

- EN is the weight of each element node in a block

-W is the total weight of all element nodes in a block

3.5. Block Identification

Based on the entropy value obtained from the EEM, the weight of each block can be identified such as nosy block or main block. The system only extracts the block which is the maximum entropy value in a DOM tree as the informative block and also regards the remaining blocks are useless blocks in a web page.

4. Evaluation

To evaluate the system Myanmar News pages are applied. Most of the Myanmar web sites such as news sites contain significant amount of irrelevant information such as header, footer, directory lists, navigation bars, advertisements and user comments. For extracting the informative block from this information, the system measures the entropy value of the each block in a web page.

In this paper, the proposed system is tested with political, business, education, science & tech Web pages from Myanmar News sites. The Web page layout in each category is the same. The number of main block per page and the correctness of the classify block are shown in Table 1. To evaluate the accuracy of the system, precision and recall are calculated.

Precision (P) is the number of correctly classified main blocks is divided by the total number of blocks that are classified as main blocks.

Recall (R) mean the number of correctly classified main blocks is divided by the total number of actual main blocks in the test set.

News categories	Number of Documents	Number of Main Block/ per page	Correctly Classify Block
Business	10	1	1
Education	10	2	2
Entertain -ment	10	3	2
Political	10	2	2
Yatanarpone Web Portal	10	3	3

Table 1. Categories of Myanmar News Pages

According to the table 1 the precision of extracting the main block is from 77 % to around 80 % and recall is from 75% to around 85 %, respectively.

In a web page, if the heading text and body text are divided separately, the system only extracts the maximum entropy weight block. The maximum weight block is an informative block in the web pages.

5. Conclusion

In this paper, we have applied the layout of Myanmar Language Web pages to implement the Informative Block Extraction System which includes preprocessing, entropy evaluation model and block identification. The proposed system could extract the informative block based on the entropy value of a node. We normalized entropy-based mechanism to mine the informative blocks. The highest weight (entropy value) block is an informative block in a web page. When the system is evaluated, the precision is over 77% and the recall is to around

85 %. If the heading and body text of main content are blocked separately in a Web page, the system could not extract the heading of the informative block such as Entertainment Category Web pages. In the future, we will implement other remaining Myanmar Language Web sites.

References

[1] C. Li, J. Dong, and J. Chen, "Extraction of Informative Blocks from Web Pages Based on VIPS", Journal of Computational Infromation Systems6:1(2010) 271-277.

[2] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma, "VIPS: a Vision- based Page Segmentation Algorithm", Technical Report, MSR-TR, Nov. 1, 2003.

[3] L. Yi, B. Liu, and X. Li, "Eliminating Noisy Information in Web Pages for Data Mining", in Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2003).

[4] P. M. Joshi, and S. Liu, " Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing", ACM, DocEng, 2009.

[5] P. S. Hiremath, S. S. Benchalli, S. P. Algur, and R. V. Udapudi, "Mining Data Regions from Web Pages", International Conference on Management of Data COMAD, India, December 2005.

[6] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm,"DOM-based Content Extraction of HTML Documents", *Pro.* 12 th International Conference on WWW, ISBN: 1-58113-680-3, 2003.

[7] S. H. Lin and J. M .Ho, "Discovering Informative Content Blocks from Web Documents", in Pro. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.588-593, July 2002.

[8] T. Htwe, "Cleaning Various Noise Patterns in Web Pages for Web Data Extraction", International Journal of Network and Mobile Technologies, VOL 1/ ISSUE 2/ November 2010.

[9] T. Win and K.N.N.Tun, " Noise Elimination for Improving Web Information Extraction", in Pro, International Conference on Computer Application (ICCA), Feburary 2009.

[10] Y. Fu, D. Yang, and S. Tang,"Using XPath to Discover Informative Content Blocks of Web Pages", IEEE. DOI 10.1109/SKG, 2007.

[11] Y. Li and J. Yang, "A Novel Method to Extract Informative Blocks from Web Pages", IEEE. DOI 10. 1109/ *JCAI*, 2009.

[12] http://www.w3.org/DOM

[13] http://www.tidy.sourceforge.net