Web Page Categorization Based on Content and Data Extraction for Academic Community

Khaing Wah Wah Linn, Sabai Phyu
University of Computer Studies, Yangon
khaingwah2linn@gmail.com, sabaiphyu72@gmail.com

Abstract

The web is a large amount of data and difficult to search information or data of user interest (IT academic field). Therefore, it needs to categorize for meet user's interesting field easily. Web page categorization help improve the quality of web search. In this paper, we proposed a framework for web data extraction by using categorized web pages to improve data extraction accuracy and result. Firstly, the numbers of test web pages are defined as inputs. We use page segmentation algorithm (VIPS) to perform segmentation these pages to achieve content structure for web page cleaning and to evaluate informative or main content block. These main contents are categorized by using Support Vector Machine (SVM) which gives accurate and efficient result. These categorized web pages are stored into the database (IT library) to output data accurately when user query.

Keywords: VIPS, SVM, Web Page Segmentation, Categorization, Data extraction

1. Introduction

There are increasing amount of data available on the web and almost 1.5 million web pages are being added daily. Therefore, the organization of these pages can't be easily searched. Besides, classification for efficient and accurate web search is very essential. Some of the approaches are widely used in web page classification system such as K-Nearest Neighbor approach, machine-learning algorithm, Bayesian probabilistic models, Decision tree, neural networks, Support Vector Machines (SVM) and so on. Furthermore, a web page usually contains the number of noise which is not related to the main information of this page such as navigation bar, advertisements, related articles and so on. Noise on the web pages tends to problem mining the main content of these pages. Generally, web page designers would gather the content of the web page to make it easy for reading. Thus, semantically related content is usually grouped together and the entire page is divided into regions for different contents. The html web page contains lots of data, but the information that the user wants is called main information (main content on the web page) and the rest are noises. Many researcher research on extraction of information and categorization from the web pages in different domain but this paper deal with literature domain on IT Academic web sites.

In our system, section 2 highlights the related work. We illustrate theory background of the system and approach in section 3. In section 4, we discuss framework of our system. Finally, this paper shows the conclusion and future work in section 5.

2. Related Work

The major approach focuses on web page categorization for web data extraction". Many researchers have undergone infinite researches for removing noisy data from the web pages. Tag tree or Document Object Model provides each web page a tree structure. In [14] Lin & Ho, proposed a method, to discover informative content blocks from web documents. In [9] Swe Swe Nyein proposed a system based on the CST tree generated by the DOM tree and also could extract the relevant documents from the web pages using cosine similarity measure. A Vision-based Page Segmentation (VIPS) algorithm is proposed in [2], that segments web pages using DOM tree with a combination of human visual cues, including tag cue, color cue, size cue, and others. The VIPS algorithm has been applied to information retrieval, information extraction and learning block importance on a single html web page [6, 7, 15].

After retrieving the web page's main information, this main information is further processed to data mining technique like Classification. In [16] proposed an approach "Text categorization with support vector

machines: learning with many relevant features" which has done to categorize the information into predefined categories.

3. Method and Background Theory

VIPS (Vision-based Page Segmentation algorithm) and SVM (Support Vector Machine) used in web page categorization and extraction of proposed framework is presented as follow:

3.1. Motivation for VIPS and Web page Segmentation

Many researcher have described in comparison the number of web page segmentation method such as FixedPS, DOMPS, VIPS, and so on. In *FixedPS*, fixed –length passages are used to overcome difficulty of length normalization. A fixed length passage contains fixed number of continuous words. The main shortcoming of the fixed-length is that no semantic information is taken into account in the segmentation process. In *DOMPS*, provide each web page with a fine-grained structure, which illustrates not only the content but also the presentation of the page. This method tends to partition pages based on their predefined syntactic structure, i.e., the HTML tags. DOM is a linear structure, so visually adjacent blocks may be far from each other in the structure and departed wrongly. Moreover, tags such as <TABLE> and <P> are used not only for content presentation but also for layout structuring. Furthermore, DOM prefers more on presentation to content. Therefore, it is not accurate enough to discriminate different semantic blocks in a web page.

Vision-based Page Segmentation algorithm (VIPS) as shown in figure 1 extracts the blocks structure by using some visual cues and tag properties of the nodes. Unlike DOM-based page segmentation, a visual block can contain DOM nodes from different branches in the DOM structure with different granularities. Structural tags such as <TABLE> and <P> can be divided appropriately with the help of visual information and wrong presentation of DOM structure can be reorganized to a proper form. Therefore, VIPS can achieve a better content structure for the original web page. Because of these benefits, we use VIPS algorithm in our system for web page segmentation. This algorithm makes full use of page layout features and tries to partition the page at the semantic level as shown in figure 2. In this step, we output content structure as shown in figure 3. Each node in the extracted content structure will correspond to a block of coherent content in the original page [3, 6].

3.1.1 VIPS Algorithm

It has three steps: block extraction, separator detection and content structure construction. These three steps as a whole are regarded as a round. Extracting visual blocks consists of a top-down tree through the DOM tree. This step is recursive and in each iteration a new node representing visual block is detected in the DOM tree. After this detection a decision is made based on certain properties of blocks (color, size,...) whether the block shall be recursively segmented further or not . Visual blocks are represented by *vb* in figure 2. It illustrates the layout structure and the vision-based content structure of the page. The original web page has four objects or visual blocks *vb1~vb4*.

The second step is separator detection as shown in figure 2. Separator is represented by μ. Separator is defined as horizontal or vertical line or rather rectangular area which doesn't intersect any of previously detected blocks [6]. Separators among these blocks are identified and the weight of a separator is set based on properties of its neighboring blocks.

The final step of VIPS is content structure construction as shown in figure 3. In this step we iterate through a list of previously found separators and merge visual blocks adjacent to them. It's important to merge blocks adjacent to separators with the smallest weight first. Before merging we have to check whether blocks meet granularity requirement. If they do, there is no need for merging them. The granularity requirement is represented by PDoC and the general rule for meeting the granularity requirement is that DoC> PDoC. If the block doesn't meet the requirement, we return to step one with root node being that visual block as shown is figure 1.

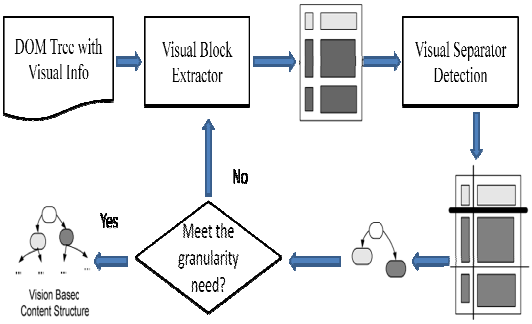

Figure 1. Steps of VIPS algorithm

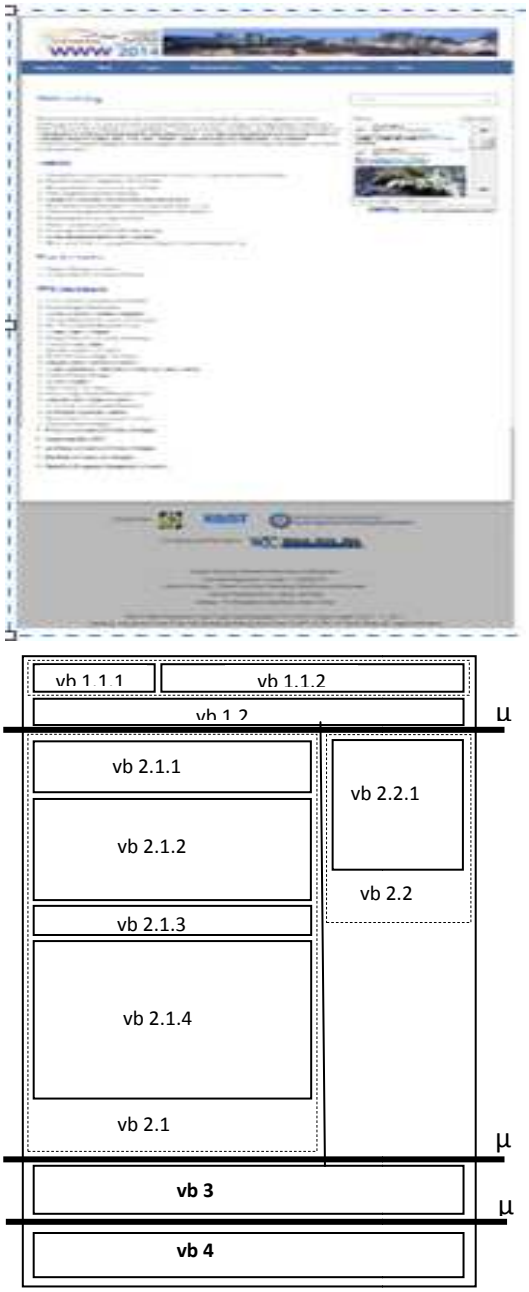

Figure 2.(a) IT field web page layout structure and (b) vision based content structure of an example page

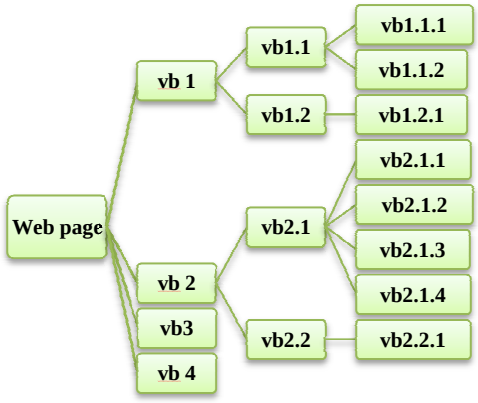

Figure 3. Vision based content structure

3.2. Introduction to Support Vector Machine

A support vector machine (SVM) is one type of learning system, which has many desirable qualities that make it one of most popular algorithms. It not only has a solid theoretical foundation, but also performs classification more accurately than most other algorithms in many applications. Especially, those applications involve very high dimensional data. It has been shown by several researchers that SVM is perhaps the most accurate algorithm for text classification. It is also widely used in Web page classification and bioinformatics applications [10].

3.3. Cosine Similarity

There are many similarity measures. The most well known one is the *cosine similarity* which is the cosine of the angle between the query vector q and the document vector d_j. We estimate the similarity of content in the web pages using this measure. Then we calculate the weight of each node (term) in content structure. We use TF-IDF (Term Frequency and Inverse Document Frequency)[10] to calculate the weight of each node.

eqn.1

w_{ij}= term weight in the document

w_{iq}= term weight in the query

After weight calculation, we estimate the similarity of these nodes using their weights. Then, we eliminate all other nodes except the highest similarity node. Ranking of the documents is done using their similarity values. The top ranked documents are

regarded as more relevant. We will assume that the high similarity nodes are main content of web page.

4. System Architecture

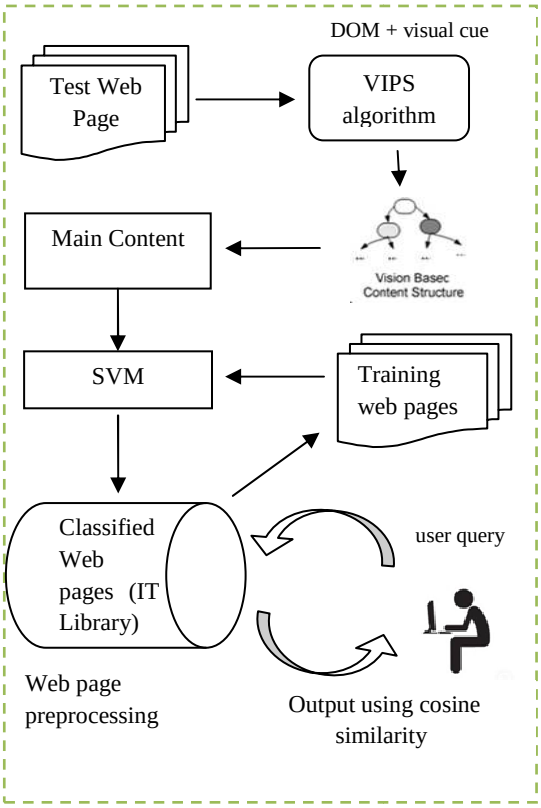

Figure 4. Web data extraction and web page categorization

We are describing our proposed system "Web Data Extraction and Web page Categorization", we build IT web pages library to response accurate web pages as shown in figure 4. For this library, we take a number of web pages and this page is segmented by VIPS algorithm that is mentioned in section 3.1.1. After this process, the number of different blocks is achieved. These content blocks are not only main content of that page but also noise such advertisement, navigation bar, and so on. Therefore, we find the similarity measure to retrieve main contents of that page. Blocks that are the high importance degrees in each web page are saved into database and eliminate the noise that are low degrees blocks. Finally, these main contents of the web pages are categorized with SVM. For example, we have taken a web page which may or may not belong to the above mentioned IT's academic field. It is showed that our experimental web page belongs to which predefined category. In final part of my framework, IT_field web pages data are extracted to the user based on their query terms. User query time can be reduced and data extraction is optimized using this framework.

For this particular system, we took 750 web pages from Information Technology education web sites, including *en.wikipedia.org*, *www.informingscience.us/icarus/* to train. These are web pages which contain not only news/ articles concerning Web Mining, Machine Learning, Cryptography, Software Engineering, Digital Image Processing but also noisy data. These web pages are segmented using VIPS algorithm and the performance of this algorithm as shown in figure 5. We compared our segmentation method with the DOM page segmentation algorithm.

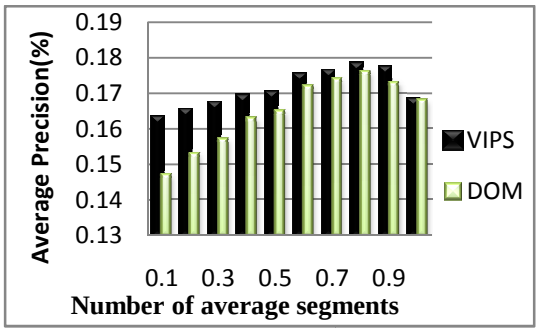

Figure 5. Comparison of VIPS and DOM

Table 1. Web Page Segmentation Accuracy for selected Web Site

URL	Precision	Recall
en.wikipedia.org	100	98
informingscience.us	98	97
computerweekly.com	92.4	87.5
Average	96.8	94.2

5. Conclusions and Future Work

We have described an approach for the data extraction by using web page classification. There are many web page categorization and data extraction approaches. Most of the approaches were proposed based on DOM tree. In this paper, web pages are segmented based on DOM tree and visual cue. Because of using visual cues, this method is more accurate than DOM tree. The produced web content structure is very useful for applications such as web adaptation, information retrieval and information extraction. Web content structure can effectively represent the semantic structure of the web page. We improve the traditional data extraction methods. When user query a IT field's term (for example_ mining), search engine can reply not only IT field's web page such as web mining, data mining, content mining and so on, but also other web pages such as

ore or metal mining and so on. To tackle this problem, we proposed this framework. This system can retrieve the most accurate data for IT field because IT_web pages have categorized. There are mainly two parts in our system. Firstly, IT library are built and then extract web data that user query. In our work, we only take IT field category of WWW and consider web document text. It could be extended to categorize the web pages into very broad categories.

References

[1] Bing Liu. Web Content Mining. The 14th International WWW Conference(WWW-2005), Chiba, Japan.

[2] Cai, D.,Yu, S., Wen, J.R., Ma, W.Y., "VIPS: A vision-based segmentation algorithm". 2003.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, " Introduction to Information Retrieval", Cambride. Uniersity Press, New York, USA, 2008, 2009.

[4] C.Chung Chaing and C.Jen Lin , "LIBSVM: A Library for Support Vector Machines",National Taiwan Universiy, Taipei, Taiwan, 2013.

[5] Elgin Akpinar and Yeliz Yesilada, "Vision Based Page Segmentation: Extended and Improved Alorithm", Middle East Technical University, Ankara, Turkey.

[6] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Extraction Content Structure for Web Pages based on Visual Representation", Microsoft Research Asia, China.

[7] Amit Chauhan, Himanshu Uniyal, Dr.Bhasker Pant, "Cleaining Web Pages for Relevant Text Extraction and Text Categorization", Graphic Era University, India.

[8] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Block-based Web Search", Microsoft Research Asia, China.

[9] Swe Swe Nyein, "Mining Contents in Web Page Using Cosine Similarity", University of Computer Studies, Yangon, Myanmar.

[10] "Web Data Mining" , Springer, page -190.

[11] Ling Liu and M. Tamer Ozsu (Eds.), "Ontology to appear in the Encyclopedia of Database System", Springer-Verlag, 2008.

[12] Arul Prakash Asirvatham, Kranthi Kumar. Ravi, "Web page Categorization based on Document Structure", INDIA.

[13] J. Alamelu Mangai, Dipti D. Kothari and V. Santhosh Kumar, "A Novel Approach for Automatic Web Page Classification using Feature Intervals", Dubai 345055, U.A.E.

[14] Lin, S.-H. And Ho, J.-M. 2002. Discovering informative content blocks from web documents. In Proceedings of the 8th ACM SIGKDD Knowledge Discovery and Data Mining, Edmonton, Canada.

[15] C. Li, J. Dong, and J. Chen, " Extraction of Informative Blocks from Web Pages Based on VIPS", January 2010.

[16] Thorsten Joachims, " Text categorization with support vector machines: learning with many relevant feature", 1997.