ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

Drawing Out Information from WebPages with Visual Unit

Mayur Raut[1], Antriksh Borkar[2]

Assistant Professor, Dept. of Electronics and Telecommunication, Sinhgad Institute of Technology, Lonavala, Pune, Maharashtra, India[1]

BE Student, Dept. of Electronics and Telecommunication, Sinhgad Institute of Technology, Lonavala, Pune, Maharashtra, India[2]

ABSTRACT: The Document Object Model is a convention to represent and interact with objects in HTML, XHTML and XML documents. It uses a tree structure to represent objects in HTML. To extract information from web pages, leaf nodes of DOM tree can be considered as basic objects. Web pages aren't made up of individual DOM tree nodes, but are more of information blocks having a visual consistent format. These blocks may not be linked to the DOM tree nodes. This paper proposes to extract the news content by visual unit which will be a visual oriented extraction method. According to visual features and text features, the visual units are identified by a top down approach. The page content is extracted according to the domain characteristic. The method is an encouraging approach for news content extraction consisting of visual units and domain characteristic. The weight parameters can also be trained by machine learning technique. Also, in the content extraction phase we can analyse that which visual unit was selected as the main content unit so that precision can be obtained.

KEYWORDS: Information extraction, DOM, Top-down approach.

I.INTRODUCTION

The ever developing Internet brings up new challenges for searching the information and its discovery. Due to the huge volume of information, finding out information of interest is becoming more and more difficult [1]. Therefore, to find the desired information, extraction of information is destined to become a useful technology. As a major part of information extraction, content extraction has provided rich corpus [2] for natural language processing, text classification, and text mining [3].

Earlier work on content extraction brought up the result that wrappers consume a lot of time because they are made for some specific source. Other, two kinds of approaches, DOM-based [2] and vision-based [4], are researched widely. Approaches that are DOM-based use text features for generating the extract rules by taking the leaf nodes of DOM tree as the smallest unit of information. However, vision-based approaches are implemented on the fact that web pages usually consist of information blocks which each have a consistent visual format [5][6] containing either one or more than one DOM tree nodes. If the information blocks are identified correctly with the usually features, then, intuitively vision-based approaches can be highly effective. The layout of some web pages can be complex and divided into various domains of application which hinders the performance of current researches. As one of the most representative vision-based approaches, VIPS [5] divides web pages into some visual blocks by heuristic rules. The various VIPS rules are defined on the basis of html tags with respect to the segment logic of general page that might not support some domains on specific application.

This paper proposes the idea of extracting content from news pages by a visual united based method. The visual units are visually 'atomic' elements which cannot be divided for smaller units on the vision. On the basis of heuristic rules like VIPS, they can be identified by a top-down approach. Taking the domain characteristics of news web pages into consideration, the definition of segment rules is independent of HTML tags. For effective work to be done on the different news websites, the extraction of the page content begins with respect to the domain characteristics of preferring some more text and fewer hyperlinks. Ranking of visual units is specifically done by the calculation of various parameter features like hyperlink density, punctuation, text density, etc. Our approach made is not dependent of HTML. Making it effective, feasible so that the universality can be improved.

The remaining structure of the paper is as follows: Related works are briefly reviewed in section II. We have provided an idea of extracting the content on the basis of visual units and evaluated its performance in Sections III and IV. Section V consists of the conclusions.

II. WORK RELATED IN BRIEF

Information extraction is the procedure of transforming unstructured or semi-structured textual data into structured representation [7]. Earlier work on content extraction is mainly by wrappers [8] which are usually a set of rules designed to extract data from a specific information source that share structural similarities [9]. There are different ways of constructing wrappers that are manual, automatic and semi-automatic approach. In [10], an interactive wrapper was proposed and it could reduce the amount of user work for training accurate wrappers through designing a suitable training interface. An unsupervised approach was explored in [11] which use Linked Data for Wrapper inducting. But, the process of generating wrappers could consume a lot of time and be error-prone when it comes to different information sources.

DOM-based approach extracts data by parsing webpage into DOM structure and searches the target nodes by several domain features [12]. DOM nodes containing main content are always found by some features such as text density, punctuation, hyperlink, etc. In the news pages, we all know that the main content usually has more text, fewer hyperlinks, more full stop '.' etc. [13]. CoreEx [14] was motivated by the observation that there was more text than hyperlinks in the main content of news [3]. The ratio of text and anchor text of DOM nodes was calculated and the one with the highest value was selected. An automatic extraction approach was proposed by ECON [15]. DOM nodes containing text were found. After that, in accordance with the continuity of DOM tree label, target node was located from these leaf nodes upward. The paper [2] believed that noise in new pages was usually highly formatted and contained more hyperlinks. In contrast to this, most of the news content has simple format containing a lot of text. Punctuation density as an important feature to extract page content was used in paper [16]. These methods stated above took leaf nodes of DOM tree as the smallest information unit to generate extraction rules using text features.

The approach based on the vision extracts page content using visual blocks coming from vision based page segmentation making it different from the DOM-based approach. Gu X D et al. [17] proposed a top-down procedure to split a web page. Dividing and merging blocks helped to detect the content extraction of web pages. Cai D et al. [5] proposed a vision-based page segmentation algorithm, VIPS. In accordance to the visual information such as page layouts, colour cues, font size etc.., the web pages is divided into visual blocks. Based on VIPS, visual blocks were represented by feature vectors in paper [18] and were merged if feature vectors were found similar. The group which contained the main content occupying the largest rectangular area, located in the middle and containing much text. Autmann Y et al. [19] demonstrated an approach based on a hierarchical structure of the visual representation locating desired data from set of documents which were manually marked training. Luo P et al. [20] put a method which extracted content on the basis of visual information and DOM tree dividing web pages into visual blocks in accordance to tag features of DOM tree.

Though, that vision-based segmentation was based partially on HTML tags, changing the structure of the web pages might make it invalid. Therefore, for improving the generality, the paper proposes an approach which is not dependent on HTML tags. The smallest visual units of DROM tree based on paper are detect and then the visual unit containing the page content in accordance to text and visual features are selected.

III. EXTRACTION OF NEWS CONTENTS WITH VISUAL UNIT

News content extraction is the process of separating the news body from other noise information [21]. Usually a news page also consists of other information than the actual news, which might be considered as noise consisting of advertisements, links to related news, copyright, etc. The presented information always has the information blocks having a recurring format. The extraction of content can become more effective if the blocks of information are identified correctly with the visual features. The paper proposes an approach which detects blocks of information with visual and text features. Domain characteristic helped to select the main content information block.

The process of visual unit identification and content extraction is described in this section.
• Visual Unit Identification.
Visual units can be defined as visually 'atomic' elements which can't be divided into further smaller units. A visual unit may be a leaf node or a sub-tree. To judge if a DOM node is a visual unit or not, format features are used.
- The horizontal row tag <HR> is used for page layout purposes and they contain no information.
- The background colour, font of colour must be consistent because the visual unit format should be internally similar.
- The news page show us the news text containing information in terms of words with a unified style which can be viewed as a visual unit.
- The hyperlinks on the news page are not difficult to find as they are related to the current page.
The heuristic rules for judging whether to segment the node or not are produced below.

RULES:
RULE 1: Cut the node if the text length = 0 of the children of DOM node.
RULE 2: Divide the DOM node if its background colour doesn't match with one of its children's.
RULE 3: Divide this node if the colour of the font of DOM node doesn't match with its children's.
RULE 4: Don't divide if child nodes of DOM node are text nodes. Text node is that node whose ratio of total character length and hyperlink length is more than a threshold number.
RULE5: Don't divide if child nodes of current DOM node are link nodes. Link node is that node whose anchor text length is more than the text length.

In RULE 4 and RULE 5 are based on text features which do not depend on HTML tag. It helps to increase the generality of extraction system. The visual units are detected by the top down procedure as following:

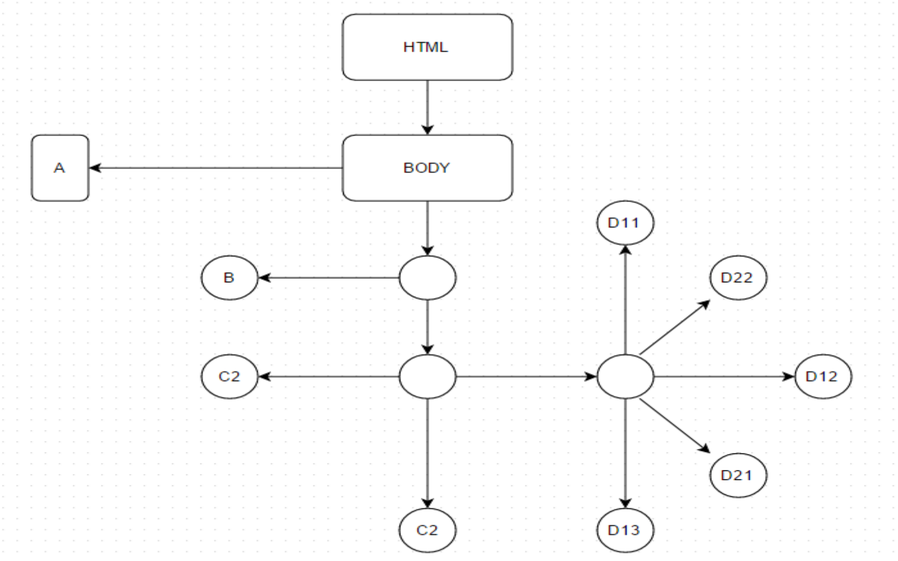


Fig. 1: DOM Tree

DOI: 10.15662/IJAREEIE.2015.041020

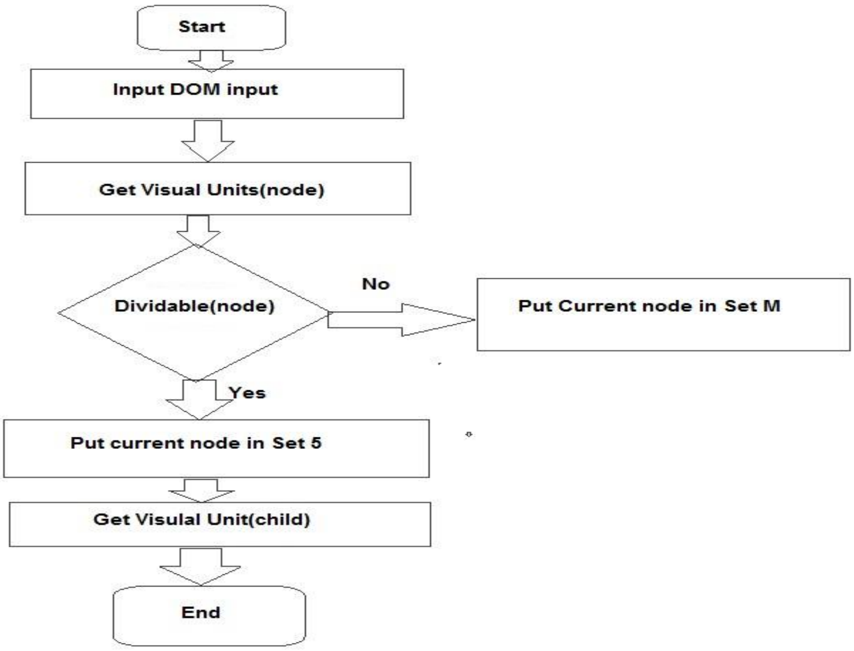


Fig.2: Flowchart for Visual Unit Detection

In Fig.2, flowchart for Visual Unit detection is been shown. It depends on the Rules given above. Output will be Visual Unit if node is unavailable.Fig.3 shows the algorithm for node segmentation.

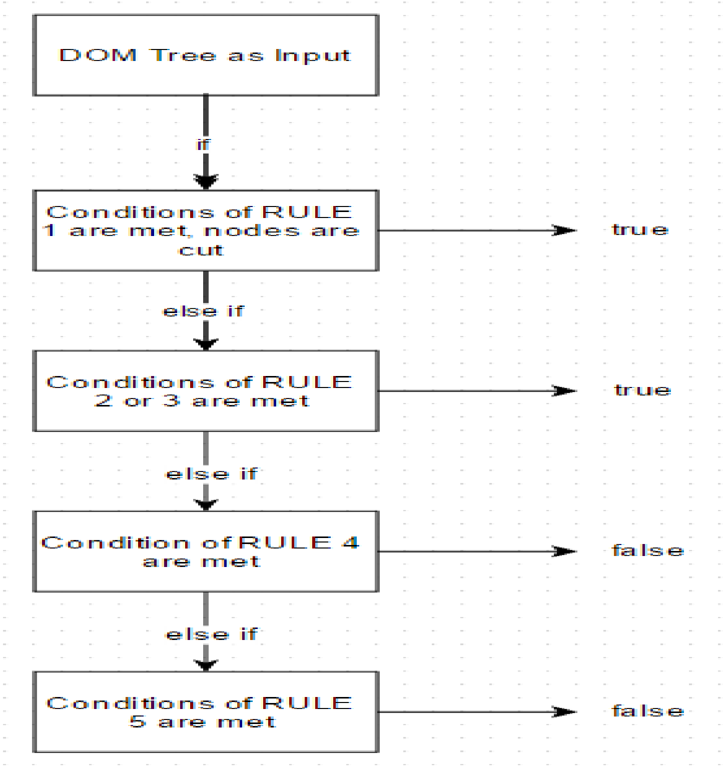

DOM Tree as Input
if
Conditions of RULE 1 are met, nodes are cut
true
else if
Conditions of RULE 2 or 3 are met
true
else if
Condition of RULE 4 are met
false
else if
Conditions of RULE 5 are met
false

Fig. 3: Algorithm for node segmentation

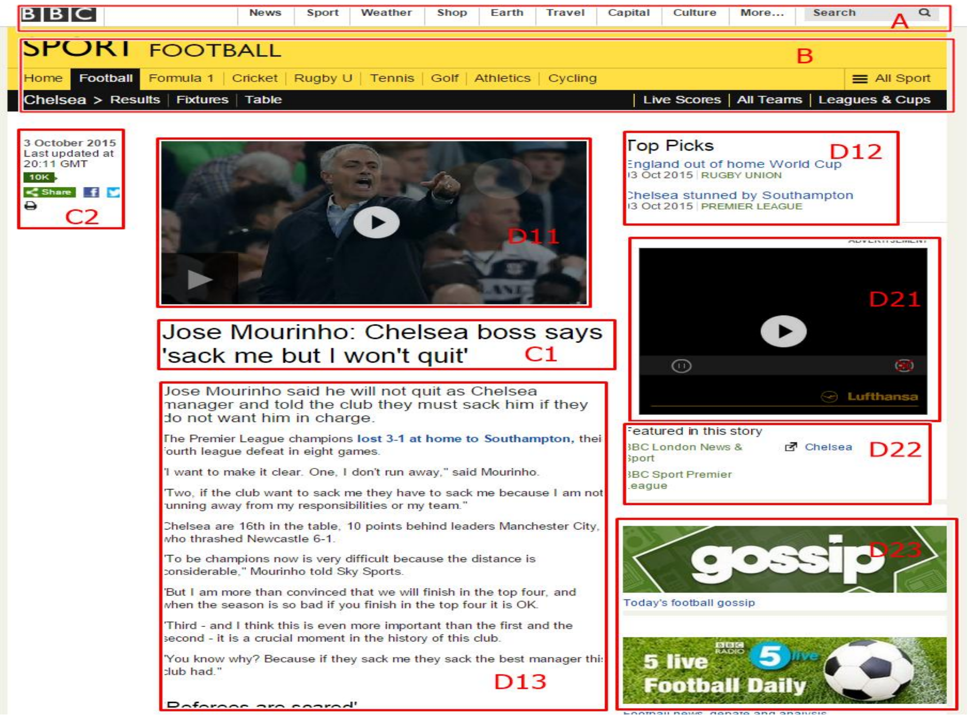
ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

To explain the above process in Fig.4 with an example, we take a news page from BBC. Considering the Fig.1, our body ‘node’ is divided into two further nodes ‘A’ and the other nodes due to satisfying the condition of RULE 2. A set 5 is made and the children of the node ‘body’ are put into it. The process keeps continuing even after the set 5 is empty. The page is segmented into ‘A’, ‘B’, ‘C1’, ‘C2’, ‘D11’, etc. The Fig. 1 is a partial DOM tree for the page. Though, the leaf nodes in Fig. 1 may not be consistent with the DOM tree of page as shown in Fig.4. In Fig. 4, the marked unit C1 is the visual unit which is the leaf node of the page and the rest of the visual units are sub-tree. The other visual units can be obtained by the above procedure.



Jose Mourinho: Chelsea boss says 'sack me but I won't quit'

Jose Mourinho said he will not quit as Chelsea manager and told the club they must sack him if they do not want him in charge.

The Premier League champions lost 3-1 at home to Southampton, thei fourth league defeat in eight games.

"I want to make it clear. One, I don't run away," said Mourinho.

"Two, if the club want to sack me they have to sack me because I am not running away from my responsibilities or my team."

Chelsea are 16th in the table, 10 points behind leaders Manchester City, who thrashed Newcastle 6-1.

"To be champions now is very difficult because the distance is considerable," Mourinho told Sky Sports.

"But I am more than convinced that we will finish in the top four, and when the season is so bad if you finish in the top four it is OK.

"Third - and I think this is even more important than the first and the second - it is a crucial moment in the history of this club.

"You know why? Because if they sack me they sack the best manager thi: club had."

'Referees are scared'

Fig.4: News page on which experiment is performed

CONTENT EXTRACTION:

The task of content extraction is to search a visual unit containing the main content from the result of visual unit identification phase. In any news page, the information of the main content provides the visitor the hint of the news. It contains larger text compared to other areas as well as less hyperlink and more full stops. Whereas the features of the noise information are different it occupies lesser area. The conclusion as a result, features visual units which are provided below.

- Area Feature:

It is defined as the ratio of visual unit area and page area.

$$S_a = \frac{A(vu)}{A(page)} \tag{1}$$

S_a is the ratio of visual unit area and page area. $A(vu)$ Stands for geometric area of visual unit and $A(page)$ stands for the geometric area of page. Larger area is mainly occupied by the main content of the page.

- If a full stop ‘.’ is found in the visual unit, S_p is set to 1 otherwise 0.
- The ratio of hyperlink text length and total text length of visual unit is defined as the link text density feature.

$$S_{td} = \frac{L(Link\ Text)}{L(Text)} \tag{2}$$

$L(Text)$ is the length of text in visual unit and $L(Link\ Text)$ is the length of anchor text. S_{td} is the ratio of hyperlink text length and total text length of visual unit. The main content in the news pages has fewer hyperlinks. Hence, S_{td} is inversely proportional to the probability of the visual unit contains the main content.

- Text Density Feature: It is the ratio of no-hyperlink text length and total text length

$$S_{td} = \frac{L(NoHyperlink\ \ Text)}{L(Text)} \qquad (3)$$

S_{td} is the value of visual unit with text density feature. $L(NoHyperlink\ Text)$ is the no-hyperlink text length of visual unit. If the value of S_{td} is more there is a higher possibility of visual unit containing page content.

- Text Length Feature: It is the text length of the visual unit.

$$S_l = \frac{L(Text)-Min(L)}{Max(L)-Min(L)} \qquad (4)$$

$Min(L)$ is the minimum length of all visual units' text and $Max(L)$ is the maximum length. The main content always has more text length of visual unit than the other units.
To sum up, the total score of visual unit is:

$$Score(vu) = w_aS_a + w_pS_p - w_{ld}S_{ld} + w_{td}S_{td} + w_1S_1$$

$$w_a + w_p + w_{ld} + w_{td} + w_l = 1 \qquad (5)$$

The total score of visual unit is $Score(vu)$. The weights that contribute to the $Score(vu)$ are $w_a, w_p, w_{ld}, w_{td}, w_l$. Weights value can be adjusted to change their contribution to the total score. The process is as follows.

a. First of all, the pre-processing of the news page is done. Some tags like select, form, input, text area and option are neglected because usually they do not carry any valuable content. Script nodes are also ignored because they give dynamic effects to web.
b. After the pre-processing is done, the phase of visual unit identification helps to divide the news page into several visual units.
c. The total score of each visual unit is calculated by the formula.
d. The unit having the highest score is selected as a target.

IV. EXPERIMENTS

This paper proposed an extraction method based on visual unit. The news pages are divided into visual units and they are given some score and one having the highest score is selected. We used different news pages; the HtmlParser parsed the web page into DOM structure and utilized WebBrowser controls to get visual information. Table 1 shows results of a few websites we took for experimentation. The number of pages undertaken was 50. The 'Right' column shows the number of pages which have the precise news content we needed. The 'Inexact' column shows the number of pages which have the redundant information along with the actual news content. The 'Error' column shows the phase which has error. Thus, this method is very effective and precise.

Table 1: Results

Website	Pages	Right	Inexact	Error	Accuracy
BBC	50	46	2	2	92%
SkySports	50	45	2	3	90%
Mirror	50	47	1	2	94%
Telegraph	50	48	1	1	96%
ESPN	50	45	2	3	90%

V. CONCLUSIONS

We proposed an approach of extracting news content with visual unit in this paper. We used the top down approach based on visual features and text features. The segmentation rules are independent of HTML tags in the visual unit identification phase, so that the page structure doesn't influence our approach. We also took various parameters like visual unit area, link text density, punctuation, etc. into consideration. Our approach is very promising for news content extraction with domain characteristics and visual units. The vision-based approach is in accordance with the human habit of web page browsing. Machine learning technique can be used to train the weight parameters. Visual unit can be selected as the main content unit for content extraction to obtain precision.

REFERENCES

1. He Y, Qiu M, Jin M, et al. Improvement on HITS Algorithm[J]. APPLIED MATHEMATICS & INFORMATION SCIENCES, 2012, 6(3): 1075-1085.
2. Sun F, Song D, Liao L. Dom based content extraction via text density[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 245-254.
3. Zhou C, Chen H, Tao J. GRAPH: A Domain Ontology-driven Semantic Graph Auto Extraction System[J]. Applied Mathematics & Information Sciences, 2011, 5(2).
4. Cai D, Yu S, Wen J R, et al. Extracting content structure for web pages based on visual representation[M]//Web Technologies and Applications. Springer Berlin Heidelberg, 2003: 406-417.
5. Cai D, Yu S, Wen J R, et al. VIPS: A vision-based page segmentation algorithm[R]. Microsoft technical report, MSR-TR- 2003-79, 2003.
6. Chen J, Xiao K. Perception-oriented online news extraction[C]//Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2008: 363-366.
7. Sarawagi S. Information extraction[J]. Foundations and trends in databases, 2008, 1(3): 261-377.
8. Kushmerick N. Wrapper induction for information extraction[D].University of Washington, 1997.
9. Irmak U, Suel T. Interactive wrapper generation with minimal user effort[C]//Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 553-563.
10. Gentile A L, Zhang Z, Augenstein I, et al. Unsupervised wrapper induction using linked data[C]//Proceedings of the seventh international conference on Knowledge capture. ACM, 2013: 41- 48.
11. Irmak U, Suel T. Interactive wrapper generation with minimal user effort[C]//Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 553-563.
12. Sun C, Guan Y. A Statistical Approach for Content Extraction from Web Page [J][J]. Journal of Chinese Information Processing, 2004, 5: 002.
13. Reis D D C, Golgher P B, Silva A S, et al. Automatic web news extraction using tree edit distance[C]//Proceedings of the 13th international conference on World Wide Web. ACM, 2004: 502- 511.
14. Prasad J, Paepcke A. Coreex: content extraction from online news articles[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 1391- 1392.
15. Guo Y, Tang H, Song L, et al. ECON: An Approach to Extract Content from Web News Page[C]//Web Conference (APWEB), 2010 12th International Asia-Pacific. IEEE, 2010: 314-320.
16. Gunasundari R, Karthikeyan S. AStudy OF CONTENT EXTRACTION FROM WEB PAGES BASED ON LINKS[J] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 2012, 2.
17. Gu X D, Chen J, Ma W Y, et al. Visual based content understanding towards web adaptation[C]//Adaptive Hypermedia and Adaptive Web-Based Systems. Springer Berlin Heidelberg, 2002: 164-173.
18. Li C, Dong J, Chen J. Extraction of informative blocks from Web pages based on VIPS[J]. Journal of Computational Information Systems, 2010, 6(1): 271-277.
19. Aumann Y, Feldman R, Liberzon Y, et al. Visual information extraction[J]. Knowledge and Information Systems, 2006, 10(1): 1-15.
20. Luo P, Fan J, Liu S, et al. Web article extraction for web printing: a DOM+ visual based approach[C]//Proceedings of the 9th ACM symposium on Document engineering. ACM, 2009: 66-69.
21. Wang J, He X, Wang C, et al. News article extraction with template-independent wrapper[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 1085- 1086.