# Design and Implementation of Hidden based Web Retrieval using Innovative Vision-based Segmentation

Kopal Maheshwari
M.E. Student, IES, IPS, Academy, Indore, India

Namrata Tapaswi
Associate Professor IES, IPS, Academy, Indore, India

## ABSTRACT
We assimilate the extracted information from a conference website to acquire the clean and high superiority academic data. This research has subsequent contributors: We propose a novel vision-based page segmentation algorithm, which use DOM tree to compensate the information loss of classical vision-based segmentation algorithm. We transform the conference Web material extraction which is difficult into a classification problematic, and categorize text blocks as predefined sets permitting to vision, key disputes, text and content information. We improve the classification quality by post-processing. Our experimental results on real-world datasets shows that our method is highly effective and efficient for extracting academic information from conference pages.

## 1. INTRODUCTION
Existing structural or semantic academic data such as ArnetMiner academic scientist social network [1] is based on databases like DBLP and ACM library. These academic data mostly defines paper publication information, but does not deliberate the academic motion knowledge. Academic conferences websites would not merely encompass paper information, but corresponding comprise much academic activity information, which contains exploration topic, conference time, location, contributors, awards, and so on. Finding such information is not merely valuable for predicting research trends, but also is the important supplement to current academic data. In order to automatically and efficiently obtain the clean and high quality academic data, it is necessary to extract useful academic information from these conference pages. Although academic conferences Web pages usually have strict layout and content description, there is no fixed extracting template for all conference pages to follow. Moreover, dissimilar sites would use different script languages to extant content. It increases the exertion of extracting information from these pages.

Similar numerous traditional extraction approaches such as extract information from new pages or deep Web pages, the information extraction from academic conference pages similar requirements a singulars solution.

In this research, we progress our preceding work [1][2][3][4] by suggesting a new technique to extract valuable academic information from conference Web pages. Major, specified illustration conference Web page, it is segmented into a established of text blocks using an algorithm which associations improve novel vision-based segmentation methods. Subsequent, text blocks are categorized into predefined categories using innovative vision-based page segmentation, in which every text block is characterized by some features those contain vision features and semantic features. Third, post-processing can expand preliminary classification consequences by repairing wrong outcomes and adding unclear outcomes. Lastly, we assimilate the extracted information from a conference website to acquire the clean and high superiority academic data. This research has subsequent contributors: We propose a novel vision-based page segmentation algorithm, which use DOM problematic, and categorize text blocks as predefined sets permitting to vision, key disputes, text and content information. We expand the classification superiority by post-processing. Our experimental outcomes on real-world datasets demonstrated that our technique is extremely effective and active for extracting speculative information from conference pages.

## 2. RELATED WORK
Various methods for mining information from Web pages have been anticipated [5, 6]. This paper distributes these conventional information extraction mechanisms into four classifications permitting to the automation level: Manually built IE. The comparison with 2011 rewe assimilates the extracted information from a conference website to acquire the clean and high superiority academic data. This research has subsequent contributors. For manually-constructed IE systems, general programming languages such as Perl or special-designed languages are used to contribution users to design a wrapper for all Website by hand. Such definitive systems comprise TSIMMIS [7], Minerva [8], Web-OQL [9]W4F [10] and XWRAP [11]. Typically, these systems have low effectiveness and are not mountable for large scale extracting tasks. Supervised IE systems yield the labeled Web pages as illustration data and then output a wrapper. In such systems, users can be trained to label the data instead of programmers, thus it reduces the cost of wrapper generation.

CaiDongdong in et al [1] this paper analysis and forecast the after-sale demand of products. They have selected fuzzy analytic hierarchy process to evaluate the quality of after-sale services and establish an evaluation indicator system by combining concrete cases. The system pays attention to the selection of the after-sale quality in the point of the strength about the enterprise competition. The accuracy of the evaluation system and method is verified through examples.

RadhouaneBoughammoura in et al [2] they have proposed three main contributions:, A new model for query representation: this model, provides matching between elements of query and elements of query interface. A new approach of query interpretation and extraction: they have proposed approach emulates capacity of users to interpret query interfaces, to evaluate method on two standard datasets.

Huilan Zhao in et al [3] in this paper, an automatic classification algorithm of Deep Web sources based on iterative self-organizing data analysis techniques algorithm according to query interface characteristics is presented.

Conventionally, retrieving data from hidden web sites has two tasks: resource discovery and content extraction [5]. The first

task deals with the automatically finding the relevant Web sites containing the hidden information. The second task deals with obtaining the information from those sites by filling out forms with relevant keywords. The original work on hidden Web crawler design [5] focused on extracting content from searchable databases. They introduced an operation model of HiWe (Hidden Web crawler). A useful observations and implications are discussed about hidden Web in [4]. They give a clear observation that the crawler strategy for deep web are likely to be different from surface web crawlers. In [8], form focused crawler for hidden web is described which utilizes various classifiers to extract relevant forms. An adaptive strategy based crawler design is discussed in [3]. The paper is relevant to the previous approaches. By using the above works a prototype system is developed for resource discovery and content extraction.

ViDE [7] uses a visual based approach for aligning data records. Unlike existing automatic wrappers, data items are differentiated using their visual properties rather than DOM Tree structure. Using the size, relative and absolute positions of data items, ViDE wrapper is able to align data records based on their size and position in the web page. Data items which are similar in size are grouped and categorized and the priority of alignment is given to data items which are located on top and to the left of the data items under consideration.

## 3. PROPOSED METHODOLOGY

Every these preprocessing technique repeatedly have a regular aspect and that is page segmentation. Its undertaking is to separate particular page to minor blocks which are consistent whichever logically or visually, based on input parameters and used algorithm. Essential segmentation technique can be dividing into two groups: DOM-based (text-based) and vision-based. Techniques in the previous group is based on evaluating a web page exclusively for several requirements to represent it. That means choosen methods are moreover based on examine HTML code straight or navigate the DOM tree matching to the HTML code and estimate information assemble from it. Distinction and speed of this method is regularly entirely based on used heuristics. The array of heuristics can be dissimilar from untainted text estimate [1] to multipart algorithms enchanting a extensive selection of possessions into account. Though these technique constantly fail to obtain one extremely important feature into explanation and that is layout of the page. As Radhouane converse in [3], the DOM base replica isn't precisely recounting real relation of creature blocks in expressions of their visual manifestation. If the complexity of CSS is in use into account, any node of DOM tree can be located at a absolutely dissimilar component of a page when evaluate with the situation in the DOM tree. It can be still undetectable, thus almost missing. DOM-based technique in the literature is in common a great deal closer than vision-based technique and caching their consequences nearly all probable wouldn't acquire greatly performance expand. As formerly expose the reason for their speed is that they don't compute all the information contained in CSS about the true layout of the inspected page. Therefore in further text we would be interested only in vision-based segmentation methods. This family of methods is based in an approach with a simple concept but quite large computing demands. The concept is to identify blocks on a web page as any user would perceive them if he was looking at the rendered page in his browser. This implies an advantage of these methods over DOM-based and it is not being strictly limited to web page processing but also being applicable to PDF and other document formats.

Table 1. Class Table

| **Class** | **Class Content** | **Explanation** |
|---|---|---|
| Date | Data Item | Conference Event |
| Location | Place Item | Address |
| Research | Topics | Research Area |
| People | People Item | People, Name and Institution |
| Paper | Paper | Paper Type and Author Title |

Vision-based segmentation algorithms have to simulate user's view of given web page, which means a page has to be rendered either to an actual picture or at least to a corresponding internal representation of the visual information contained on that page. This process of rendering is very complex due to complexity of both HTML and CSS specifications. That means demands both for computational power and time to process one page is quite high, which is problematic. After being rendered, the page has to be segmented in several iterations which is also very demanding. The most commonly used algorithm in the area of vision base segmentation is VIPS and algorithms using it as a black box and improving its results. Another approach, partially derived from the original VIPS specification, has been offered by Burget[1]. To extract the academic information, we first segment Web pages into blocks by VIPS [3], which is a popular vision-based page segmentation algorithm. VIPS uses page structures and some vision features, such as background color, text font, text size and distance between text blocks, to segment a page. Although VIPS can obtain good segmentation results for most pages, it will also lose important information for some pages. The reason is that VIPS algorithm is mainly based on vision features of page elements, therefore, the display of segmented adjacent semantic blocks may be same and then VIPS would ignore the blocks whose display is inconsistent. Therefore, we introduce DOM-based analysis to improve the VIPS segmentation results, namely, we will find these missed text blocks.

Since some blocks such as navigation, copyright and advertisement blocks do not contains the academic information. We regard these blocks as noise, which should be removed from VIPS complete tree. The noise removing process uses some vision features [4]. Position features comprise block position in horizontal and vertical on page and ratio of block area to page area. Layout features contain alignment of blocks, whether neighbor blocks are overlapped or adjacent. Appearance features include size font, image size, and font of link. Content features consist of common words of blocks and particular order of some words. According to these vision features, we can remove noise nodes from VIPS absolute tree.

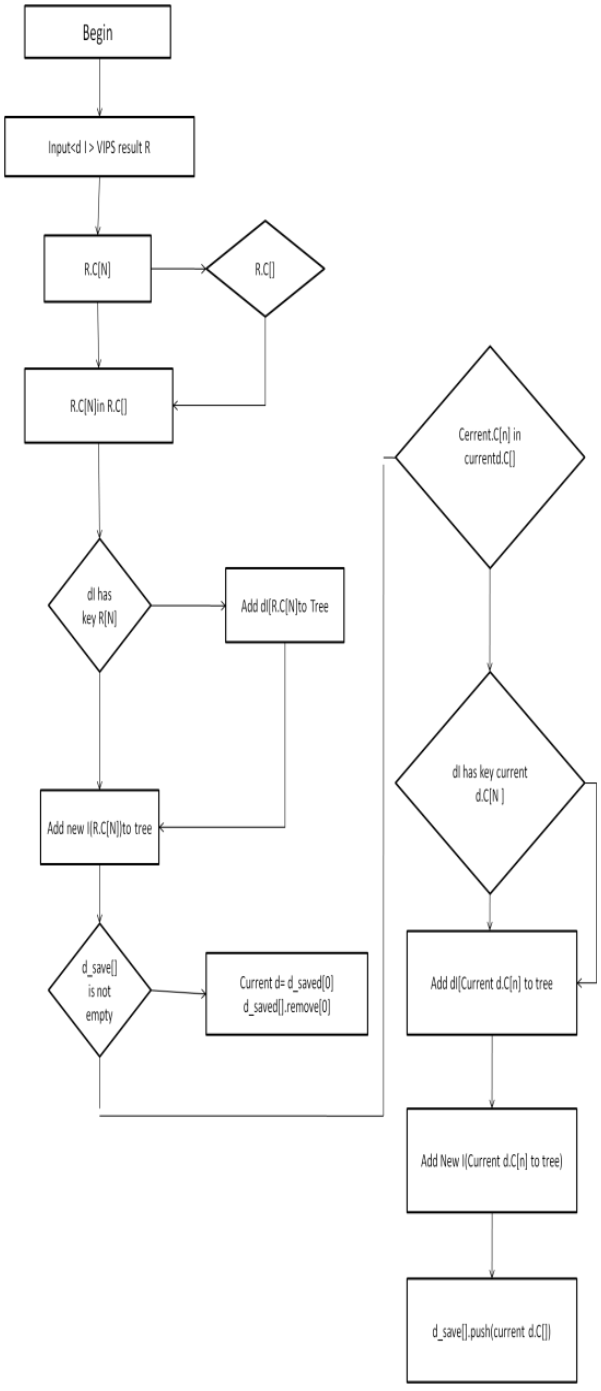


**Figure 1: IVIPS (innovative vision-based page segmentation)**

**Table 2. Feature Vectors of Text Blocks**

| Vector | Description | Value |
|---|---|---|
| isHeader | Whether the biggest font size | true, false |
| isTitle | Whether the title font size | true, false |
| nearestTitle | Type of the nearest title block | int |
| textLength | Length of text block | int |
| fontSizeToAverage | Average font size | int |
| fontWeightToAverage | Average font weight | int |
| startWithLi | Whether start with <li> | true, false |
| dateTypeNum | Number of key words about date type, such as *deadline* | int |
| dateNum | Number of key words about date, such as *January* | int |
| placeTypeNum | Number of key words about location type, such as *Place* | int |
| placeNum | Number of key words about location, such as *Italy* | int |
| areaNum | Number of key words about research area | int |
| nameNum | Number of names | int |
| institutionNum | Number of institutions | int |
| positionNum | Number of positions | int |
| authorNum | Number of authors | int |
| abstractTypeNum | Number of key words about abstract | int |
| paperTypeNum | Number of key words about paper type | int |
| wordNum | Number of words of text blocks | int |
| wordToName | Ratio of number of words to number of names | double |
| linkTotext | Ratio of length of link to length of blocks | double |
| left | Ratio of left margin to page width | double |
| width | Ratio of block width to page width | double |

We can select some features to measure a given text blocks. Therefore, we use some vectors as shows to describe each block. For a text block, we construct its feature vectors according to vision, key words and text content information. For example, given a text block: "Paper Submission Due: Monday, dec 12, 2013 (23:59 UTC -11)" and its HTML source code "<li> Paper Submission Due:  Monday, dec 12, 2013 (23:59 UTC -11)</li>", its feature can be constructed as:

(1) Vision features: isTitle=false, isHeader =false, startWithLi=true, left=(280-0)/950=0.3 (page width:950, left margin: 0, text left margin:280), with=640/950=0.7 (text width: 640);

(2) Key word features: nearest Title=DI (its nearest and isTitle=true blocks is about date information), dateNum=2 (it contains 2 date words), paperTypeNum=1 (it contains 1 key word about paper: Paper);

(3) Text content features: fontSize=0, fontWeight=0, textLength=58, textLink=0, wordNum=11, nameNum=5, wordToName=11/5=2.2.
There are many famous existing classification algorithm such as C4.5, K-Nearest Neighbors and Bayesian Network. C4.5 and Bayesian Network are the most widely used classifier models. In our previous work [2], we have showed that these classification algorithms will produce very similar results, and innovative vision-based page segmentation model performs little better than other models. Therefore, we still choose Bayesian Network model to solve the text blocks classifier problem.

# 4. PROPOSED ALGORITHM

SB = vision semantic block(b)
LN =design node(d) of vision tree generated by VIPS

DN = information node(i)

Result = R

we need to obtain the basic vision semantic blocks by analyzing DOM tree of Web pages. A vision semantic block is a text block with independent meaning, and it is between two newline tags such as <br/> and only contains style tags and texts. A lot of blank nodes are removed from HTML tags. Then we traverse DOM tree to extract vision semantic blocks.

Algorithm 1 shows the detail of generating the VIPS complete tree. Let SB be the vision semantic block, LN be the layout node of vision tree generated by VIPS, and DN be data node.

This algorithm includes three steps:

(1) It finds a SB by traverse LN to search matched layout nodes;

(2) If it finds a matched layout node, then this node is also a vision semantic block;

(3) If it does not find a matched layout node, then add this SB into the vision tree.

This algorithm not only assures that there is no information loss, but also preserves the structure of the vision tree.

Input: <LayoutNode,DataNode> LNIN[], VIPS result PN

Output: a VIPS complete tree T

begin for (PN.children[i] in PN.children[])

if (LNIN[] has key PN.children[i])

add LNIN[PN.children[i]] to T

else

add new DataNode(PN.children[i]) to T

LN_saved[].add(PN.children[i])

end

while (LN_saved[] is not empty) {

currentLN = LN_saved[0]

LN_saved[].remove(0)

currentDN = LNIN[currentLN]

for (currentLN.children[i] in currentLN.children[])

if (LNIN[] has key currentLN.children[i])

add LNIN[currentLN.children[i]] to T

else

add new DataNode(currentLN.children[i]) to T

LN_saved[].push(currentLN.children[i])

end

end

end

# 5. EXPERIMENT RESULTS

(1) Due to the heterogeneity of different conference Web pages, some rule-based Web information extraction techniques are not scalable any more. The rules extracted from one conference Web site can not apply to another conference, so we should find out an approach independent from page templates.
(2) A lot of existing IE systems uses a DOM tree to represent HTML page and complete information extraction based on the structure of the DOM tree. But HTML tags does not follow strict grammar restrict, it is likely to cause an error in parsing HTML DOM tree. In addition, DOM tree is initially designed to display data in the browser, rather than describe the semantic structure of Web pages, so even though two nodes have the same parent node in the DOM tree, it does not mean they are more closely in semantic than other nodes.

(3) Traditional information extraction systems always take a single Web page as input, but the useful information of a conference can be located in multiple pages of the Web site, so the system must perform information extraction from Web site level, and integrate the extraction results of each page to complete information extraction. First experiment is verifying the complete tree. the Web page segmentation module is implemented in C#.net an open-source machine learning library, to classify text blocks.

Our experimental results are obtained on a PC with 1.99GHz CPU, 2GB RAM and Windows 8. We can see that the complete trees have more leaf nodes than vision trees. It means our algorithm can find more text blocks than VIPS.

**We observed some facts:**
(1) There are many noise blocks in the complete tree. In some websites, almost half of all blocks are noise blocks.

(2) Our removing noise method can remove average 39% noise nodes and 51% noise leaf nodes. Therefore, it will reduce the number of nodesshould be processed in extraction and improve the efficiency. The third experiment is the comparison between initial classification results and the results after post processing. The results are obtained on 20 randomly websites. We have two conclusions:

(1) The initial classification results only have average 0.75 precision, 0.67 recalls and 0.68 F1- measures. After post-processing, the classification results are improved to average 0.96 precision, 0.98 recall and 0.97 F1- measure. Therefore, the post-processing key roles in academic information extraction.

(2) Some text blocks like DI, PO, PE and TO, which have clear vision and text content features, have better classification results. The average F1-measure on these blocks is 0.99.

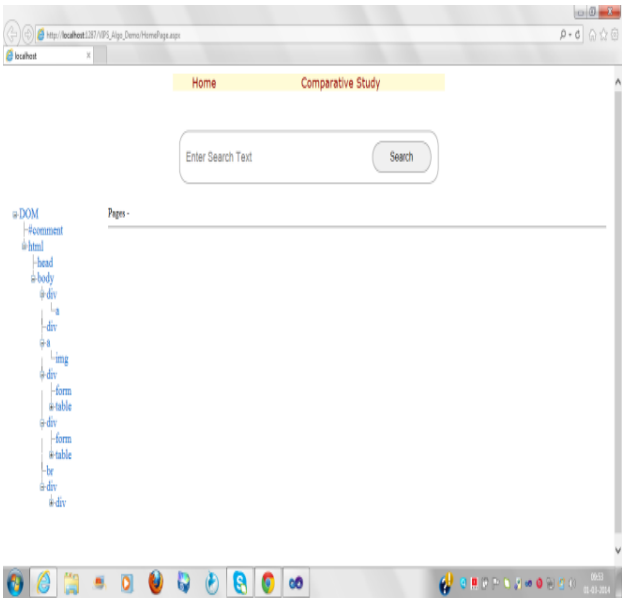


**Figure 2: Implementation of Hidden Web Retrieval**

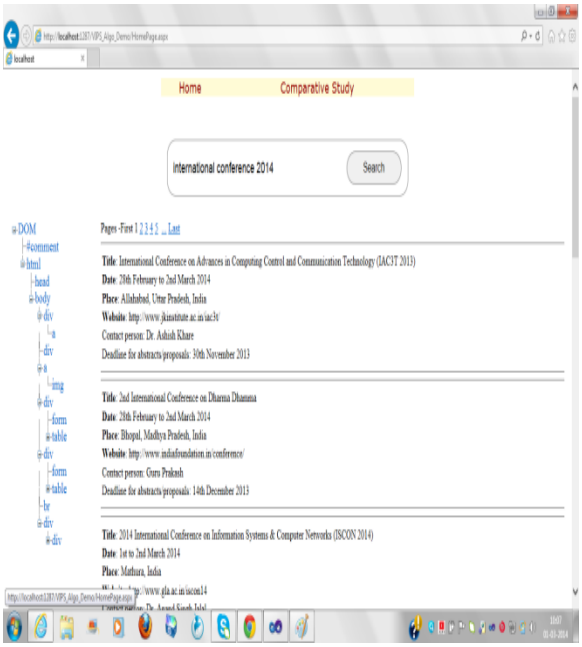


**Figure: 3 Result of Hidden Web Retrieval**

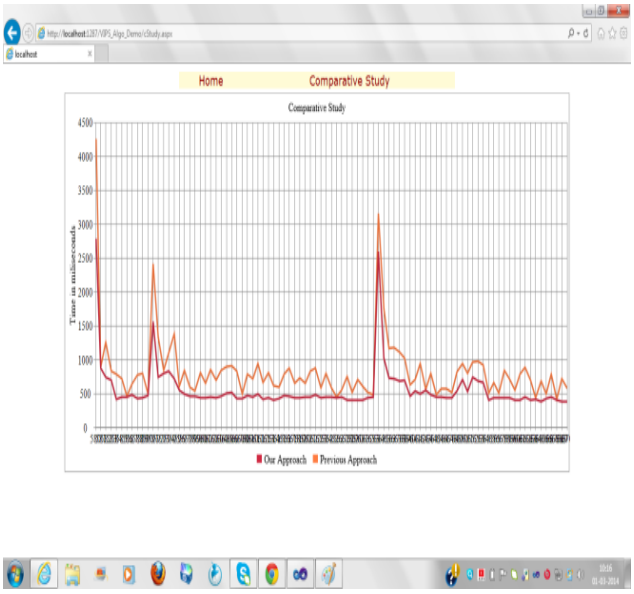


**Figure: 4 Comparison of Hidden Web Retrieval**

Finally, we compare our extracted results with previous work. Figure 4 compares results of new method with previous work [2].

It can be seen that our new method improves the 2011 result greatly: the precision is improved from 0.90 to 0.96, the recall is improved from 0.89 to 0.98, and the F1 is improved from 0.90 to 0.97. The reasons of our improvement are: (1) We have designed a new algorithm to segment pages, and this algorithm not only finds missed blocks, but also removes some noise blocks; (2) We propose the more reasonable categories to classify text blocks; (3) More post-processing rules are used in our new method.

## 6. CONCLUSION

This research proposed a novel technique to extract functional educational information from conference Web pages automatically. Primarily, given an example of conference Web page, it is segmented into a set of text blocks with an algorithm which combine vision-based segmentation method and DOM-based segmentation method. Subsequently, text blocks are classified into pre-defined category and post processing on the initial classification consequences can improve the classification. At last, we combine the extracted information from a conference website to obtain the clean and high quality academic data.

## 7. REFERENCES

[1] CaiDongdong, CaiDongdong, Zhang Tianrui and Wang Xiao 2012 , "Research of After-sales Service Management System Based on Web" International Conference on System Science and Engineering June 30-July 2, 2012, Dalian, China.

[2] Radhouane Boughammoura and LobnaHlaoua, Mohamed NazihOmri, "VIQI: A New Approach for Visual Interpretation of Deep Web Query Interfaces", Computing Technology and Information Management (ICCM), 8th International Conference on (Volume: 1) 24-26 April 2012.

[3] Raghavan, S. and Garcia-Molina, H. 2001, "Crawling the Hidden Web", VLDB Conference presentation 129 – 138.

[4] Rekha Jain and Dr. G. N. Purohit Department of Computer Science, Apaji Institute, Banasthali University, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications (0975 - 8887) Volume 13- No.5, January 2011.

[5] Wei Liu, XiaofengMeng and WeiyiMeng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions On Knowledge And Data Engineering", VOL. 22, 2010.

[6] Jayant Madhavan, David Ko, ŁucjaKot, VigneshGanapathy, Alex Rasmussen and Alon Halevy. "Google's DeepWeb Crawl". PVLDB '08, August 23-28, 2008, Auckland, New Zealand.

[7] Gang Liu, Kai Liu, Yuan-yuan Dang, "Research on discovering Deep web entries Based ontopic crawling and ontology" 978-1-4244-8165-1/11 IEEE -2011.

[8] Chelsea Hicks, Matthew Scheffer, Anne H.H. Ngu and Quan Z. Sheng", Discovery and Cataloging of Deep Web Sources" IEEE IRI 2012, August 8-10, 2012.

[9] Zilu Cui and Yuchen Fu, "Deep Web Data Source Classification Based On Query Interface Context", Fourth International Conference on Computational and Information Sciences- 2012.

[10] Dayne Freitag, "Information extraction from HTML: Application ofageneral learning approach," Proceedings of the 15th ConferenceonArtificial Intelligence (AAAI1998), Madison, Wisconsin, USA, 1998.

[11] Mary Elaine Califf and Raymond J. Mooney, "Relational learning of pattern-match rules for information extraction," Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Stanford, California, USA, 1998.

[12] Peng Wang, Yue You, Baowen Xu, and Jianyu Zhao, "Extracting Academic Information from Conference Web Pages," The 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Boca Raton, FL, 2011.

[13] Stephen Soderland, "Learning information extraction rules for semistructuredand free text," Journal of Machine Learning, vol. 34, pp. 233-272, 1999.

[14] Nicholas Kushmerick and Daniel S. Weld, "Wrapper induction for information extraction", Proceedings of the 15th International Conference on Artificial Intelligence (IJCAI1997), Nagoya, Aichi, Japan,1997.

[15] Ion Muslea, Steve Minton and Craig Knob lock, "A hierarchical approach to wrapper induction", Proceedings of the 3rd International

[16] Conference on Autonomous Agents, Seattle, Washington, USA, 1999