A Semantic DOM Approach for Webpage Information Extraction

Yulian Fei
Computer Science and Information Engineering Institute,
Zhejiang Gongshang University,
Hangzhou, China
fyl@mail.zjgsu.edu.cn

Zongwei Luo
E-Business Technology Institute,
University of Hong Kong,
Hongkong, China
zwluo@eti.hku.hk

Yun Xu
Computer Science and Information Engineering Institute,
Zhejiang Gongshang University,
Hangzhou, China
xuyun@mail.zjgsu.edu.cn

Winston Zhang
University of Hong Kong,
Hongkong, China
winstonzhang@lscm.hk

Abstract: **With the development of electronic technology and e-commerce, technology for web pages has attracted a lot of research efforts which becomes one of the hottest topics recently. This paper has proposed a semantic DOM(SDOM) approach for information extraction of e-commerce WebPages. With the combination of content and structure information, the precision and recall can achieve a good result which is shown in our experiments on listpage and tablepage data sets.**

Keywords: semi-structured; information extraction; SDOM

Ⅰ INTRODUCTION

With the development of electronic technology and e-commerce, semi-structured information extraction technology for web pages has attracted a lot of research efforts which becomes one of the hottest topics recently. Many approaches are adopted in this specific research domain such as machine learning [1], data mining [2] and conceptual modeling [3] etc. to obtain process-required information. Differs from the characteristics of the general WebPages, bulletin board systems, blog, facebook and professional B2C websites resemble each other very much in the sense that the informative data are stored in a tabular format such as table, list, div and etc. Those formats can be repeated many times even in a single page to form a common topic. This kind of repetition and concentration requires new research efforts to acquire better extraction performance.

Name of Product	Spec	Material	Producing Area	Price (RMB/ton)	Up/Down	Remark
Twisted Steel	φ10mm	HRB335	ShaGang	3510	-	
Twisted Steel	φ10mm	HRB335	JiYuan	3460	-	
Twisted Steel	φ10mm	HRB335	PingGang	3460	-	
Twisted Steel	φ12mm	HRB335	ShaGang	3420	-	
Twisted Steel	φ12mm	HRB335	MaGang	3380	-	
Twisted Steel	φ12mm	HRB335	YongGang	3380	-	
Twisted Steel	φ12mm	HRB335	PingGang	3370	-	
Twisted Steel	φ14-16mm	HRB335	ShaGang	3370	-	
Twisted Steel	φ18-25mm	HRB335	ShaGang	3380	-	
Twisted Steel	φ16-25mm	HRB335	MaGang	3380	-	
Twisted Steel	φ16-25mm	HRB335	PingGang	3350	-	

Figure1: standardized table webpag

There exists many general information extraction techniques [4] [5] [6] [7] [8] [9] [10] [11] [12], seldom of them are customerized for this newly emerging domain. The most frequent patterns appeared in this domain is regular planar table, shown in figure 1 and 2, which could be table, list, div and etc. Each row can be seen as a new data entry under a common category. Each column (if exists) is an attribute of this entry. But the structure of this kind of pages does not contain the corresponding semantic meanings. Also the WebPages when created might be inappropriate, such as some missing values or irrelevant entries. To achieve a fast and accurate extraction in this domain, it is naturally to combine the semantic meaning of the content with its corresponding structure information. Yet, the current techniques pay few attentions on this approach. To our best knowledge, this paper is the first work to introduce the concept of semantic document object model (SDOM) to deal with this new domain.

Name: 120*1190-700*1310
Production Area:__
Description: Resource Number 622363 variety Profile manufacturer GudiZhe material: 1.2738model Tax-inclusive price 13000 RBM
Vender: XinShunLuTeGang
See detailed information

Name: High Quality Q235 Flat Steel
Production Area:__
Description: Supply High Quality Flat steel
Remark: Plus 600RMB/ton if need to plat zinc
Vender: Yu'an machin part LLC of AnYang
See detailed information

Name: Spring Flat Steel
Production Area:__
Description: High Quality Spring Flat Steel
Remark:__
Vender:Sales Corp of TongHua Steel Group
See detailed information

Figure 2：List webpage(steel)

Ⅱ DOCUMENT OBJECT MODEL

Document object model (DOM) [13] describes a document using a tree structure. Each node in the tree manifests an HTML tag or the text item contained in the HTML tag. This kind of tree structure precisely describes the correlation between and among tags and text items in the HTML document. Such correlation includes child type, parent type and sibling type. It can make use of standard interface provided by DOM to realize the operation to nodes, including adding node, deleting node, and obtaining parent node or child node of the current

978-1-4244-4639-1/09/$25.00 ©2009 IEEE

```
<TABLE border=0>
<TR bgColor=#ffffff>
<TD>name</TD>
<TD>material</TD>
    <TD>Specifications</TD>
<TD>Price</TD>
<TD>Origin</TD>
</TR>
<TR bgColor=#ffffff>
<TD>General Wire</TD>
<TD>Q235</TD>
<TD>Φ 6.5/8mm</TD>
<TD>3550</TD>
<TD>Dangang</TD>
</TR>
</TABLE>
```

(a)

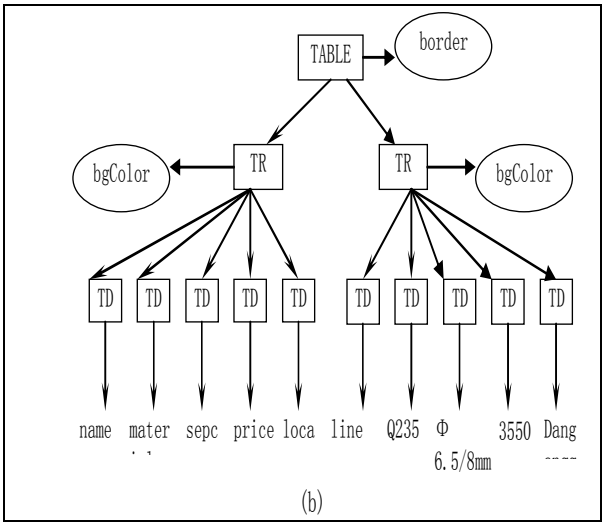

Figure3 the DOM of HTML

(active) node, etc. By the parsing of DOM, the HTML document shown in the Figure3 (a) forms the DOM tree structure which is shown in the Figure3 (b).

In Figure3 (b), rectangle denotes tag nodes in the document, ellipse represents node property, and the unframed leaf node denotes text node. According to DOM, webpage is stored in the tree structure. System-extracted Information to be extracted is some particular nodes of the tree which can be obtained by searching DOM to locate the positions of such nodes in the tree.

The webpage created through server-side dynamic technology, such as JSP, ASP, PHP, etc., essentially is static webpage with HTML tags embedded. So this kind of dynamic WebPages can also be processed by DOM.

III SEMANTIC DOM

Without loss of generality, DOM contains not semantic information but the structure of a webpage. To fast and accurate extract the useful information from massive WebPages, the semantic DOM (SDOM) is proposed in this section. To combine the conventional DOM with semantic meaning, the original DOM is mapped to the new SDOM with the help of a set of predefined semantic rules or terminologies which can be derived from some specific domain knowledge, such as the product information and business flow provided from an e-Business consultant. How to acquire diverse semantic rules is not the focus of this paper and this proposed SDOM is a generic one to incorporate diverse semantic rules if needed. To illustrate using the previous DOM tree shown in Figure3, the useful information is stored in the format of table. Traditionally, the DOM tree is then extracted as in Figure3. In this proposed SDOM, the tag <table> is replace by the semantic content contained in this table as <Product Info>, then the node <TH> is defined as <attributes> according to its illustrative content. Nodes of <TD> contained within <TH> is the detailed attributes, and each is replaced by its semantic meaning as <name>, <material> etc. The main structure of SDOM is already extracted now. The following rows of the table act as different data entries in a database, and they are converted in a row-wise manner into this new SDOM repeatedly. This converted example of SDOM tree is shown in Figure4.

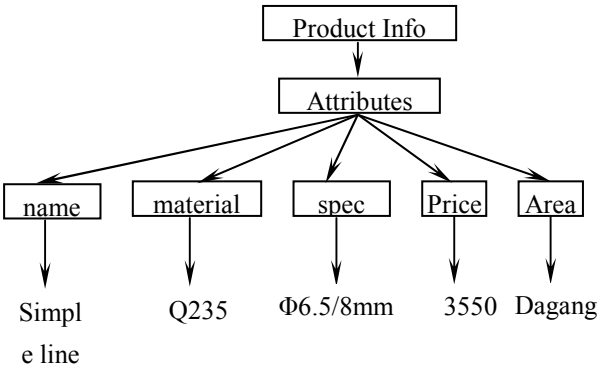

Figure4 Semantic DOM tree

IV CONSTRUCT SDOM TREE

Before the SDOM tree is constructed, we briefly introduced how the traditional DOM tree is built.

A. Construction of DOM tree

The process of constructing DOM is to decode the target page constituted under the W3C to acquire a W3C-defined DOM object. And the process is as follows:

I. To identify all HTML tags, convert codes and begin to further parse from root node <HTML>.

II. To objectify the attribute of every node.

III. To conjunct all nodes based on hierarchical structure described in the original HTML text to construct the DOM Tree.

The parser used in this paper is the source project provided by HTML standard designer W3C [14]. And the main steps of document parsing are as follows:

Process (source); // Page to be processed, provided by the search engine

```
InputStream    bis = null; //Convert codes
bis new
java.io.ByteArrayInputStream(source.getBytes("UTF-8")
);
if (bis = =null)
return;
Document dom = tidy.parseDOM(bis,null);
//Construct the document of DOM Tree
```

B. Construction of SDOM tree

For the construction of the SDOM tree, two different methods are proposed in this paper. The direct construction approach which simply replaces the DOM nodes with the semantic content contained as the illustrative example does in the Section3. The advantage of this approach lies in the quickness but it meets the problem when deals with the situation that information should be integrated together.

Therefore, in this paper we take the second approach constructing the SDOM tree through a model learning process supervised by a set of semantic rules. Main steps of this model learning algorithm can be summarized as follows:

Step 1: Identify the useful information block.

In e-commerce websites, the informative contents are usually embedded in a well organized structure like table, list etc. Therefore, directly locating those tag pairs will help increase the overall performance speed of this approach. To locate the tag pair, one can build up a key word repository reserving the tag pairs defined under the W3c standard.

Step 2: Check the machine readability of information block.

Not like Figure1, it is not a straightforward format for auto extraction in Figure2. The information block is labeled for further processing if its machine readability is poor.

Step 3: Represent the nodes.

For the cases shown in Figure3 (a), one can easily convert the DOM nodes into SDOM nodes through extracting the corresponding contents. Once blocks shown in Figure2 are found, an independent natural language processing (NLP) module is called to extract the contents and convert them into a tree structure.

Step 4: Match with the semantic rules.

In Figure5, page DB is the webpage database provided by CoSE system; DOM Parser is the parser of the webpage; Page Classifier is the classifier of the webpage which classifies the webpage to process (Tablepage, Listpage, etc.); Wrapper Adapter provides the extraction rules for the classified WebPages; SDOM module maps and convert the content into SDOM tree; and ProductInfo DB is the product database.

Once the SDOM tree is constructed, we prune the tree with the given semantic rules or terminologies with the help of NLP module. The synonyms are matched to the closest term provided in the semantic terminologies. Note that this approach itself does not provide a logic reasoning engine and the current approach can easily be extended to incorporate the third party modules to perform the reasoning process needed for complicate semantic rules.

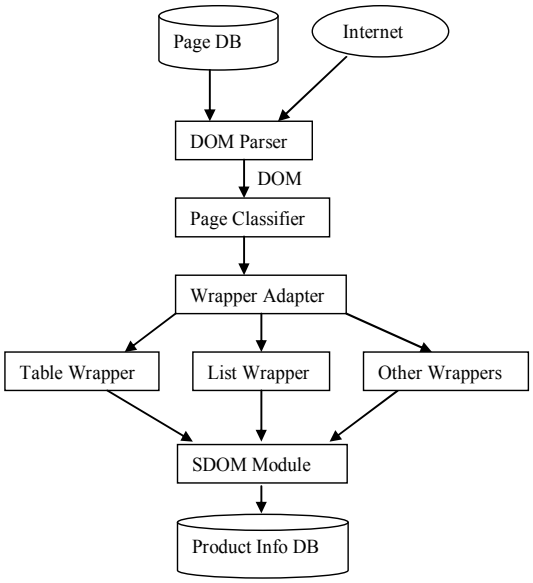

Figure5 The framework of SDIE

Ⅴ THE FRAMEWORK OF THE SYSTEM

To implement the proposed approach, SDIE (SDOM-based Information Extraction) system is designed in this section. It first sends a inquiry request to the specific information repository, wrappers extracts the required information from the responding pages, then map them into corresponding tag information[15], and SDOM module convert the tag information into SDOM tree. The CoSE Search Engine [16] serves as responding pages and extracts the related information tuple from the corresponding pages. A tuple can be defined as a vector, Si=< A1，A2…，An>, with the attribute value of n character strings. Si denotes the ith tuple, Aj denotes the jth character string. The component of SDIE includes DOM Parser, Page Classifier, constructing wrapper and SDOM module, specifically, as shown in Figure5

Ⅵ EXPERIMENTAL RESULTS

To evaluate the performance of our proposed approach, we carry on the experiments on tablepage and listpage data sets. WebPages of the steels collected is chose as the tablepage set. WebPages collected from several major e-commerce websites in China are chose as the listpage set. We extract name, specification, material, price and other detailed information (such as manufacturers, origin, contacts etc.) for tablepage, and extract name, price,

digital zoom and other detailed information. The target attribute is defined in our semantic terminologies. Results are reported in table 1 and 2.

TABLE1: THE RESULT OF THE INFORMATION EXTRACTION OF THE TABLE WEBPAGE

websites	The average scale of the webpage(byte)	The average time of construction the tree (ms)	Information number of the webpage	Extracted information number	Average extraction time (ms)	Recall rate (%)	Precision rate (%)
www.bjsteel.com	19165	52	26	26	1328	100	100
www.600bxg.com	38219	94	53	50	2954	94.3	100
www.chinasteel.com.cn	15977	32	15	15	680	100	100
www.steel.hc360.com	34969	78	39	39	2484	100	100
www.ldmetals.com	60990	47	30	30	218	100	96.8
www.m188.com	34350	15	30	30	1594	100	100
www.mysteel.com	42683	31	73	69	344	94.5	100
www.sinometal.com	30904	16	20	20	1016	100	100
www.steel35.com	50918	47	20	20	375	100	97.9
www.wzsq.com	47092	31	36	36	1813	100	100
Average	37526.7	44.3	34.2	33.5	1280.6	98.9	99.5

TABLE2: THE RESULT OF THE INFORMATION EXTRACTION OF THE LIST WEBPAGE

websites	The average scale of the webpage(byte)	The average time of construction the tree (ms)	Information number of the webpage	Extracted information number	Average extraction time (ms)	Recall rate (%)	Precision rate (%)
www.ezit.com.cn	66445	547	15	15	2031	100	100
www.800it.com.cn	85561	94	25	25	1469	100	100
www.263mall.com	46675	47	18	18	578	100	94.4
www.it88.com.cn	68937	78	30	29	1515	96.7	100
www.365e.com.cn	65997	31	16	12	985	75	100
ww.8bit. com.cn	64429	47	10	10	1187	100	100
www.zgtjsm.cn	24122	15	7	7	438	100	100
www.snuol.cn	20683	16	10	10	672	100	100
www.iibrand.com	46644	31	10	10	781	100	90
http://www.16buy.com/	36151	16	8	8	485	100	87.5
shop365.com.cn	48226	47	12	12	766	100	91.7
Average	57387	88.1	14.6	14.5	991.5	97.4	96.7

From the results, it is noticed that both listpage and tablepage achieve good precision and recall rate, especially in tablepage result. Compared with listpage, the structure of tablepage is in a better organization manner than that of the listpage. In tablepage, its semantic meaning can be more easily identified than that in listpage. Also in the listpage, much more rich information are crowed in a one single cell which obstacles the extraction ability. It is also found that the scale of the webpage is not proportional to tree constructing time. This is because the webpage facilitates constructing the tree due to its comparatively simple structure.

Results also show the importance of constructing the repository, and the precision largely depends on the integrity of the field knowledge. In tablepage, the filed information of the table can be considered as the repository. Due to its completeness, the repository of tablepage also can be seen as complete. But such information of the listpage is not complete, which leads to the comparatively poor performance of information extraction.

Ⅶ CONCLUSIONS

In this paper, we proposed a semantic DOM approach for information extraction of e-commerce WebPages. This is important due to the reason that it can utilize the semantic meaning to help increase the performance of the conventional DOM tree. With the combination of content and structure information, the precision and recall can achieve a good result which is shown in our experiments on listpage and tablepage data sets. Our proposed approach can easily be extended to a more complicate semantic environment which is important today with the development of WWW to the next generation. This work is our first attempt to integrate semantic into DOM tree, we will further our research directly along this way in the future.

REFERENCE

[1]Chun-Nan Hsu , Ming-Tzung Dung , Generating Finite-State Transducers For Semi-StructureData Extraction From The WEB , Information Systems, 8:521-538, 1998

[2]B Adelberg , NoDoSE-a tool for semi-automatically extracting structured and semistructured from text documents , Proceedings of SIGMOD’98, page 283-294, 1998

[3]D W Embley, D M Campbell, A conceptual-modeling approach to extracting data from the WEB, Proceeding of the 17th International Conference on Conceptual Modeling,, Singapore, pages 78-91, 1998

[4]Nicholas Kushmerick, Daniel S. Weld, Robert Doorenbos, Wrapper induction for information extraction, IJCAI-97, 1997

[5] Nicholas Kushmerick,Wrapper induction:Efficiency and expressiveness,Artificial Intelligence(118), pages 15-68,2000

[6]D.W.Embley , Y.Jiang , Y.K.Ng , Record-boundary discovery in WEB-documents , Proc. Of the 1999 ACM SIGMOD , pages 467-468,1999

[7]D.Buttler , L.Liu , C.Pu , A Fully Automated Object Extraction System for the World Wide WEB , International Conference on Distributed Computing Systems, pages 351-370, 2001

[8]A. Sahuguet, F. Azavant, a WysiWyg WEB wrapper factory（W4F） , Proc. 8th World Wide WEB,1999

[9] Ciravegna F, Lavelli A. Learning Pinocchio: Adaptive information extraction for real world applications. Natural Language Engineering, 10 (2), page:145-165, 2004.

[10] Larson Erik J., Hughes Todd C. Relational recognition for information extraction in free text documents. AAAI Spring Symposium-Technical Report, v SS-05-01, page:144-146, 2005.

[11] Shaker Mahmoud, Ibrahim Hamidah, Mustapha Aida, Abdullah Lili Nurliyana. A framework for extracting information from semi-structured web data sources. Proceedings-3rd International Conference on Convergence and Hybrid Information Technology, ICCIT 2008, v 1, page:27-31, 2008.

[12] Labbe Nicole, Swamidoss Isabel Maya, André Nicolas, Martin Madhavi Z, Young Timothy M., Rials Timothy G. Extraction of information from laser-induced breakdown spectroscopy spectral data by multivariate analysis. Applied Optics, 47 (31), page:158-165, 2008

[13]Joe Marini. Document Object Model. Tsinghua University Press, 2003

[14]http://sourceforge.net/projects/tidy

[15]Zhihong Guo. Extraction Technology Resources Based WEB Information. Information Science, , 20(12), 1282~1284,2002

[16] Yulian Fei, Anding Zhu, Guangmin Wang. A Topic-Based And Distributed Search Engine For Business Intelligence, DCABES 2006, pages 913-917, October 2006