2010 International Conference on Intelligent Computation Technology and Automation

Web Object Block Mining Based on Tag Similarity

Rui Liu[1], Rui Xiong[2], Kun Gao[3]
State Key Lab of Software Development Environment, Beihang University,
No.37 Xueyuan Road, Haidian District, Beijing, 100191, P.R.China
{[1]liurui, [2]xiongrui, [3]gaokun}@nlsde.buaa.edu.cn

***Abstract*—Currently, a large number of Web information on the Internet is presented in structured objects. Mining object information from Web is of great importance for Web data management. This paper presents a Web object block mining method based on tag similarity. It first constructs a DOM tree for the Web page and calculates the similarity of all possible generalized nodes. Then a pruning method is used to filter the redundant information based on the features of noise data and find the Web object region. Finally the Web objects are identified in the Web object region. The experiment results show that, comparing to IEPAD, our method got a higher precision.**

Keywords-Web Object Region; Information Extraction; Tag Similarity; DOM tree; Generalized Node;

I. INTRODUCTION

A lot of information on the Internet appears in formatted structural Web page regions which are called Web object regions. Web object region contains some similar Web objects with some important information describing them. For example, there are two books in a list in Figure 1. Both of them include some information such as description, price and so on. They are similar in structure. We can extract and integrate these object information from a variety of heterogeneous websites to provide value-added services, such as comparing with products from different websites or providing meta-search service. This paper studies how to automatically detect and extract all Web objects from Web pages.

Figure 1. An object region with two Web objects

The commonly used Web object block mining methods are based on template. By observing the layout style and the source code of the sites, the methods find out some regular patterns and then establish a template for each site to identify the data records. Although these methods are of easy realization, they are time-consuming and inadaptable for the large number of data records mining task. Other methods identify objects relying on some specific HTML tags and machine learning techniques. Cai D. et al. [3] proposed a Vision-based Page Segmentation (VIPS) algorithm. The algorithm blocks Web page according to the page layout and the vision information feature. It is a novel idea to use some heuristic factors for blocking and so achieves a good blocking effect, but it does not calculate the location and size information, so it is unable to tell whether a block is a topic region or a noise region. Chang, C-H et al. [4] proposed a method named Information Extraction Based on Pattern Discovery (IEPAD). The method uses the tag uniformity feature of Web objects to construct Patricia tree (PAT). By comparing tag sequences, it finds out a series of patterns of which each one corresponds to a Web object. The limitation of the method is that it can only extract Web objects with the same tags.

This paper provides a new method based on tag similarity called WOBM. It first constructs a DOM tree for the Web page and then calculates the similarity of all adjacent generalized nodes. According to the similarity, all Web object regions can be found and so the Web objects are extracted. We test and evaluate WOBM method by comparing with IEPAD in experiment. WOBM achieves a recall ratio of 95.94% and precision ratio of 94.50%. The result is consistent and satisfying.

II. WOBM

This paper proposed a Web object block mining method. As shown in Figure 2, it has the following steps: Web page preprocessing, constructing DOM tree, calculating similarity of generalized nodes, searching Web object region, pruning redundant information, and identifying Web object.

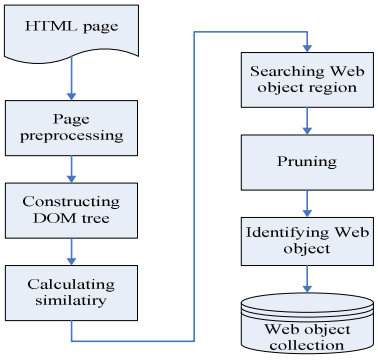

Figure 2. WOBM steps

978-0-7695-4077-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICTA.2010.684

Next we will interpret the chief steps in detail.

A. Constructing DOM Tree

Generally, Web pages are a kind of hyper text file which consists of texts, tags and others. Web page can be converted into a tag tree object according to the HTML tag structure. We call the tag tree object DOM tree.

Before DOM tree construction, we need some preprocessing work on Web pages: 1) adding extra closing tags for some tags like LI and HR. 2) deleting useless tags such as note tags and SCRIPT tags and removing them.

It is easy to construct DOM Tree based on relative theory. Currently, there are many good tools to construct DOM tree for Web page. So, we will not discuss it further.

B. Calculating Similarity of Generalized Nodes

First, let us introduce the concept *Generalized Node*.

Def. 1 *Generalized Node*: a node set composed of nodes in the HTML tag tree with the following two features:

- sharing the same parent node
- being adjacent

The length of generalized node is defined by the number of tag nodes. In Figure 3, we can find two generalized nodes. The first one is composed of the left four TR under the TABLE and the second is composed of the right four. The lengths of both are four.

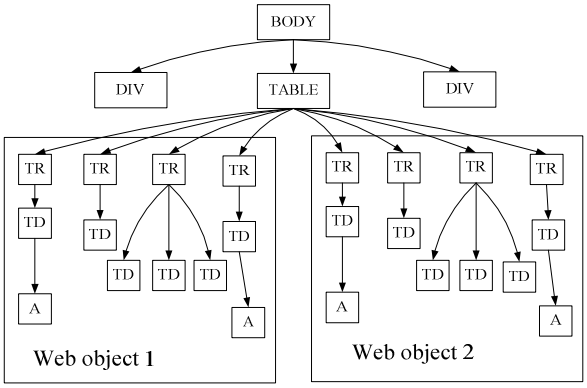

Figure 3. HTML page tag tree

Generalized node is different from tag node. We define generalized node to catch the case that a Web object may be contained in a series of adjacent nodes rather than one node. A Web object is contained in a generalized node. They are of one to one relation.

Similarity of generalized nodes is calculated through combination and comparison. Because the length of generalized nodes is unknown, it try to combines one, or two…, K (given maximum value) tag nodes as the generalized node and compare the tag similarity of them. According to statistics, the length of generalized node is small, usually less than 3 and mostly equaling 1 [4], which make the combination feasible. In order to search the location of Web object region, the comparison should start from each node in order. Note that generalized nodes share one common parent node, so the nodes to be compared should be under the same node.

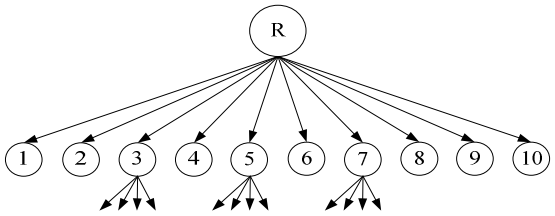

Figure 4. Combination and comparison

Take Figure 4 to illustrate, all the nodes from 1 to 10 have the same parent node R. Set K equals 3. The process performs as follows:

First, we start from node 1 and compute the following string comparisons.

- (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8), (8, 9), (9, 10)
- (1-2, 3-4), (3-4, 5-6), (5-6, 7-8), (7-8, 9-10)
- (1-2-3, 4-5-6), (4-5-6, 7-8-9)

(1, 2) means that a string sequence of a tree rooted as node 1 compares with the one rooted as node 2. The sequences are both composed of tags by depth-first mode. For example, in Figure 3, tag string sequences of the first two child nodes of "TABLE" are "TR TD A TR TD".

(1-2, 3-4) means that a string sequence of combined trees rooted as node 1 and node 2 compares with the one rooted as node 3 and node 4.

When we start from node 2, we only compute:

- (2-3, 4-5), (4-5, 6-7), (6-7, 8-9)
- (2-3-4, 5-6-7), (5-6-7, 8-9-10)

We need not do one node comparisons because they have been done when we started from node 1. From node 3, we only need to compute:

- (3-4-5, 6-7-8)

We do not need to start from other nodes after node 3, because all calculations have been done.

Assume that current node has n children, the time complexity will be O(nK), so it is acceptable.

The similarity calculation can not use simple string comparing method because Web object usually lacks of some fields. Here we employ edit distance algorithm [5, 6]. Edit distance between two strings s_1 and s_2 is the required minimum edit operations when converting s_1 into s_2. The edit operations include:

- Changing a letter
- Inserting a letter
- Deleting a letter

To facilitate the comparison and threshold determination, the edit distance is normalized:

$$NED(s_1, s_2) = \frac{d(s_1, s_2)}{(length(s_1) + length(s_2))} \qquad (1)$$

C. Searching Web Object Region

Def. 2 *Web Object Region*: a set of generalized nodes with the following features:

- Sharing the same parent node
- Having the same length
- Being adjacent
- Tag string similarity with neighbors less than the given threshold.

Web object region's generalized nodes have the same length, that is, generalized nodes under the same parent node have the same number of tag nodes. The tag nodes' sub-tree can be different so as to catch more types of objects.

We find out Web object region by deep traversal of DOM tree. Taking the current node as parent node, it searches all combination of similar nodes and selects the region with most nodes as the Web object region. Here only the generalized nodes with similarity limited in the given threshold are considered as similar nodes.

In the searching process, it may encounter the region overlap problem. Here "overlap" means one region is contained in another. As shown in Figure 5, the web page has eight data records. WOBM may submit each row or the whole region as a Web object region. To avoid this problem, the process should obey the following two principles:

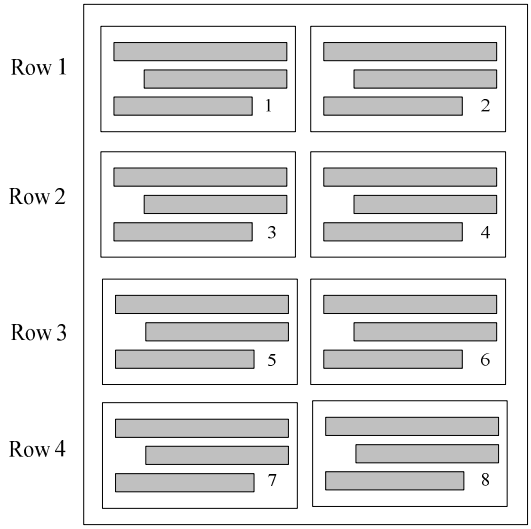

Figure 5. Overlap problem

1) Web object region overlap principle

If a high-level region overlaps a low-level one, only the former one is submitted. This principle is to avoid the situation that low-level nodes containing many small child trees are similar but not the real objects. So for the Web page in Figure 5, we submit the whole region.

2) similar string choosing principle

As we know, if a series of strings s_1, s_2, …, s_n are similar to each other, any combination of them with the same length is similar. So we only submit the generalized nodes with minimum length. In Figure 5, each row like Row 1 rather than Row 1-2 is submitted as generalized node.

D. Pruning Redundant Information

WOBM finds out the data records by tag similarity comparison. However, in most Web pages, useless information such as advertisements, navigations and classifications also have the tag similarity feature. This topic-irrelevant information in the mining results called "pseudo object". Although pseudo object harbors the tag similarity feature, it has two differences with real object:

1) The number of property fields is small. Statistic shows that pseudo objects' property fields are often less than 3, whereas real objects' are often more than 4. So we bring in an experience threshold of property field number. Object with property fields less than threshold will be pruned.

2) Link tag takes a large proportion in the whole tag sequence. We also set up an experience threshold to prune the pseudo objects.

E. Identifying Web Object

The process of Web object identification is simple. When the redundant information pruning is done, the root node will contain all Web object regions. For each Web object region, according to the records of the first generalized node's start position, generalized node's length and the total number of generalized nodes, all of the Web objects can be identified.

III. EXPERIMENTS

The main purpose of the research is to mine Web object lists, so test web pages we crawled are based on data lists like Figure 1. Here, we compare WOBM algorithm with IEPAD on identifying Web objects.

We crawl Web pages containing product lists randomly from 16 different Web sites. These pages contain restaurants, digital products, books, scientific equipment, buildings, cars and other Web objects. Table 1 shows the experimental results. First, we make a brief introduction for every column.

Column 1: It lists all sites' addresses. In this experiment we selected the 16 different sites, containing a total of 2169 Web objects. We crawl a few pages from each site randomly.

Column 2: It gives the number of Web object on each site. The object number in this column is counted by manual work.

Column 3, 4: These columns show the results of WOBM algorithm. Column 3 gives the correct number of Web objects found by WOBM and Column 4 gives the erroneous number.

Column 5, 6: These two columns are the corresponding results by IEPAD algorithm.

The last three rows show the statistical results. Recall and precision of each method is calculated based on total number of Web objects found by the methods and the real number of objects on Web pages.

As can be seen from the table, IEPAD got a good result for simple pages, such as www.verycd.com, but for the pages with complex structure and many advertisements, such as www.aibang.com, www.taobao.com and www.pconline.com.cn, the precision is low. On the other side, since WOBM calculates tag similarity by edit distance algorithm and mines Web objects with similar tag sequence

rather than the same tag sequence, it got a better result in precision for most sites.

TABLE I. EXPERIMENT RESULTS

Site	Total	WOBM		IEPAD	
		Corr.	*Err.*	*Corr.*	*Err.*
www.instrument.com.cn	127	127	0	127	34
www.gzsin.cn	120	117	1	117	21
www.scilink.cn	132	130	0	130	15
www.hbh-kytj.com	165	163	2	163	21
www.aibang.com	150	144	8	145	48
www.taobao.com	125	119	14	123	46
search.jiayuan.com	118	105	36	115	36
search.china.alibaba.com	150	142	9	148	23
car.autohome.com.cn	150	143	0	146	34
www.360buy.com	168	159	12	166	31
www.changhong.com.cn	126	123	0	123	21
tech.sina.com.cn	190	184	0	184	36
www.wl.cn/c887	128	116	12	126	26
realestate.cn.yahoo.com	120	118	0	118	13
www.verycd.com	100	96	0	96	6
www.pconline.com.cn	100	95	27	95	40
Sum.	2169	2081	121	2122	451
Recall		95.94%		97.83%	
Precision		94.50%		82.47%	
F-measure		95.21%		89.50%	

Figure 6 is the comparison chart. It shows that both of the two methods perform well on recall. But on precision, WOBM is higher than IEPAD. So the overall performance of WOBM is better than IEPAD.

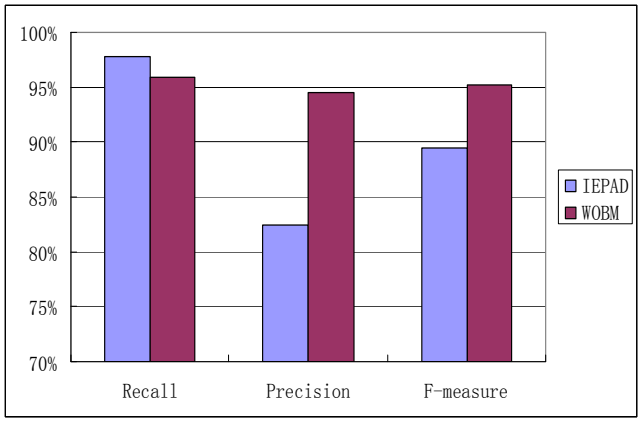

Figure 6. Comparison of IEPAD and WOBM

IV. CONCLUSIONS AND FUTRUE WORK

This paper proposed a high-precision method for automatic Web object block mining. By calculating the similarity of all adjacent generalized nodes, it finds out the Web object regions and then identifies the Web object in the regions. Experimental results show that our method gains a better overall performance than IEPAD.

Out future work will focus on the following two aspects:

1) Effective noise data clean technology. This method considers the tag similarity feature, but some useless information such as advertisements also has this similarity feature. So we will research on other technology like entropy pruning.

2) Extracting from JavaScript code. Web 2.0 makes the Web pages contain amount of JavaScript code. So next we will research on dealing with JavaScript code.

REFERENCES

[1] Robert Dale, Hermann Moisl, H. L. Somers. Handbook of Natural Language Processing [M]. CRC Press, 2000: 241-243

[2] Liu, B., Grossman, R. and Zhai, Y. "Mining data records from Web pages." KDD-03, 2003.

[3] Cai D., Yu S.P., Wen J.R. et al. VIPS: a Vision-based Page Segmentation Algorithm[R]. MSR-TR-2003-79. 2003.

[4] Chang, C-H., Lui, S-L. IEPAD: Information extraction based on pattern discovery. WWW-10, 2005.

[5] Nie, Z., Wu, F., Wen, J.-R., and Ma, W.-Y. Extracting Objects from the Web. In Proc. of ICDE. 2006.

[6] S. Soderland. Learning Information Extraction Rules for Semistructured and Free Text. Machine Learning, 1999.

[7] Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y. Simultaneous Record Detection and Attribute Labeling in web Data Extraction. In Proc. of SIGKDD, 2006.

[8] Nie, Z, Wen, J.-R and Ma, W.-Y. Object-level Vertical Search. In Proc. of CIDR. 2007

[9] Gusfield, D. Algorithms on strings, tree, and sequence, Cambridge. 1997.

[10] Buttler, D., Liu, L., Pu, C. A fully automated extraction system for the World Wide Web. IEEE ICDCS-21, 2001

[11] Chakrabarti, S. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, 2002.