Research Journal of Applied Sciences, Engineering and Technology 5(7): 2426-2432, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 26, 2012 Accepted: September 12, 2012 Published: March 11, 2013

A Grammatical Evolution Approach for Content Extraction of Electronic Commerce Website

Wei Qing-jin and Peng Jian-sheng
Department of Physics and Electronic Engineering, Hechi University, Yizhou, Guangxi, 546300, China

Abstract: Web content extraction, a problem of identifying and extracting interesting information from Web pages, plays an important role in integrating data from different sources for advanced information-based services. In this paper, an approach and techniques of extracting electronic commercial information from the Web pages without any given template is investigated in a way of Grammatical Evolution (GE) method. Although a lot of research used the Xpath technique to extract the content of Web pages, but due to the complexity of the Xpath grammar, it is too difficult to perform the processing automatically for evolutional tools. Hence, a reduced language integrating Xpath and DOM techniques is given to generate the solution of parse in a BNF grammar form, which is used in the GE. Moreover, a fitness function evaluation method is also proposed on the fuzzy membership of the two parts in the chromosome. Finally, empirical results on several real Web pages show that the new proposed technique can segment data records and extract data from them accurately, automatically and flexibly.

Keywords: DOM, grammatical evolution, web content extraction, Xpath

INTRODUCTION

Web content extraction is a problem of identifying and extracting interesting information from Web pages. It plays an important role in integrating data from different sources for advanced information-based services, such as customizable Web information gathering, comparative shopping, meta-search and so forth.

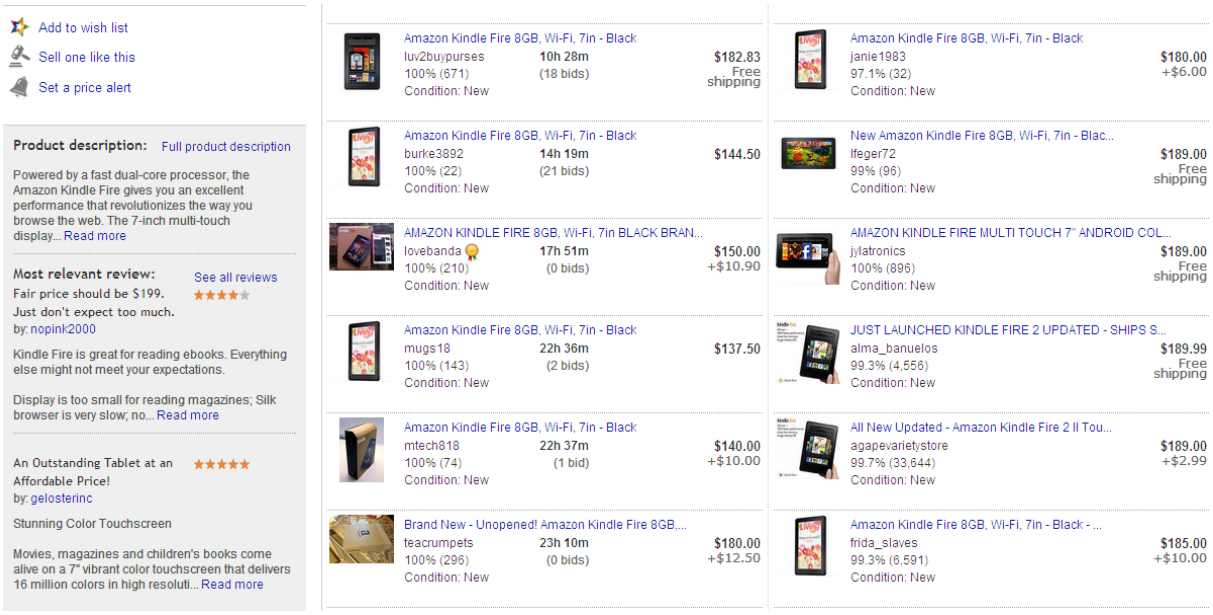
To solve this problem, several approaches ranged from heuristics, meta-heuristics, to the methods using data mining, statistics, or ontology, are used. Among them, Ziegler and Skubacz (2007) proposed an approach to extract real content from Web news pages using a particle swarm optimization algorithm. Chang *et al*. (2003) introduced a system called IEPAD to discover extraction patterns from Web pages without user-labeled examples but using several pattern discovery techniques, including PAT-trees, multiple string alignments and pattern matching algorithms. The paper (Qiu and Yang, 2010) presented a set of novel techniques based on page similarity measure, page clustering and wrapper generation to automatically extract data from E-Commerce web sites. In Reis *et al*. (2004), traditional hierarchical clustering techniques are used to extract the desired news from Web news sites. McKeown *et al*. (2003) provided an article extraction module using machine learning program Ripper.

In this study, our interests is the approach and techniques of extracting electronic commercial information from the Web pages without automatically rather than with some fixed template such as in Zhai and Liu (2007). A typical example is the product list and description pages in the electronic commerce Websites as shown in Fig. 1a and b. intuitively, these two kinds of page are slightly different in both layout and content. The former, called *list pages*, often enumerate the *list items*, i.e., a number of goods with summary introductions; but the latter, called *detail pages*, include the elaborate description of the products, such as product name, image, description and price, for comparative e-shopping.

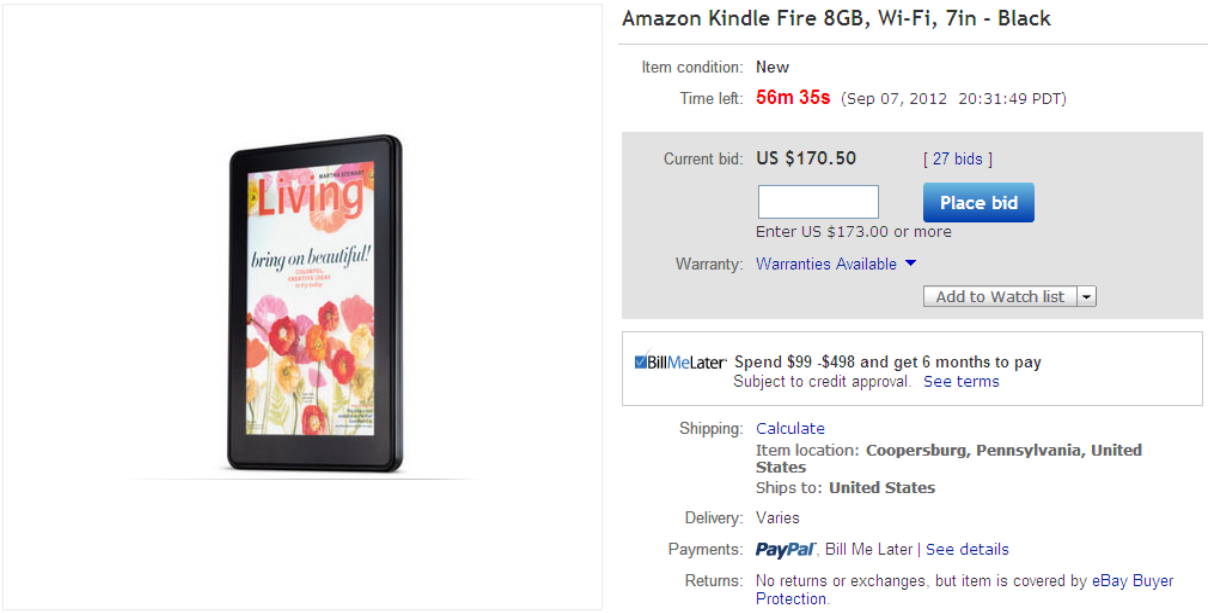
The main ideas of our attempt could be summarized as follows:

- Adopting the GE algorithm to get the expression automatically. Using this Meta heuristic method could generate the solution expression in a BNF grammatical search space rather than probes binary ones arbitrarily.
- Reducing the grammar of Xpath and combine a Dom operations into it. In this study we introduce a new language Xpath-DOM to locate the elements and attributes of the pages.
- A two segments denotation of chromosome illustrating Xpath part and DOM part also is used to feature the solution and help to calculate the fitness value.

Corresponding Author: Wei Qing-jin, Department of Physics and Electronic Engineering, Hechi University, Yizhou, Guangxi, 546300, China



(a) An example product lists page



(b) An example product detail page

Fig. 1: An example of list page and detail page in electronic commerce website

GRAMMATICAL EVOLUTION

Grammatical Evolution (GE) (O'Neill and Ryan, 2001), a variant of Genetic Programming (GP) (Koza, 1992), is an automatic programming evolutionary algorithm which includes a context free grammar and genotypes with its mapping into phenotypes. This kind of representation could select production rules in a context-free grammar in Backus-Naur form and thereby creates a phenotype. Mathematically, the grammar G is a formal grammar in which all production rules are in the form V→w where V a non-terminal symbol is and w is a sequence of terminal and non-terminal symbols. A context-free grammar can be represented by the quad-tuple: $G = (V_T, V_N, P, S)$, where V_T is a finite set of terminal symbols, V_N is a

finite set of non-terminal symbols, P denotes a set of production rules and S represents a non-terminal symbol as the Start notation.

The GE algorithm gradually replaces all non-terminal symbols with the right-hand of the selected production rule starting from the start symbol S. The substitution is defined by the following mapping Eq. (1):

$$\text{Rule} = B \bmod R_N \quad (1)$$

where,

B = A gene

R_N = The number of rules for the specific non-terminal symbol

This kind of symbol replacement process is repeated until the end of the chromosome is reached. If the final chromosome no valid expression has been produced, the algorithm repeats from the starting of the chromosome (called wrapping operation) or the mapping procedure is terminated by assigning a small fitness value to the relevant chromosome.

Due to its properties of universality, simplification and efficiency, it has been used with success in many fields such as symbolic regression (O'Neill and Ryan, 2001), Santa Fe Ant Trail (O'Neill and Ryan, 2003), discovery of trigonometric identities (Ryan *et al*., 1998), robot control (Collins and Ryan, 2000) and financial prediction (Brabazon and O'Neill, 2003).

THE PROPOSED APPROACH

In this section, an approach and techniques of extracting electronic commercial information from the Web pages without any given template is investigated in a way of Grammatical Evolution (GE) method.

Overall phrases: The schema and main processing steps of GE are the following, which are also illustrated in detail in the next sub-sections:

- **Initialization:** This step includes the set-up of the population, coefficients and relevant parameters.
- **Definition of the evolutional grammar:** In this step, a context-free grammar, describing all the possible algebraic expressions of the original set of features on both Xpath and Dom is created. A Context Free Grammar (CFG) of the Xpath-DOM Language is given in the next subsection.
- **Chromosome make-up:** Every part of each chromosome in the genetic pool is made randomly in a range of an integer interval.
- **Fitness evaluation:** Each chromosome g is evaluated in two parts: Xpath part and Dom part, which are related to Xpath feature and Dom feature, respectively. Then the fitness function could calculate the value considering the impacts on them totally.
- **Chromosome transformation:** In this phase, genetic operators, such as crossover and mutation, are imported to product the next generation of chromosomes.

Termination testing: A termination condition of the maximum number of generations or chromosome with best fitness value is tested in this step. If it reaches or exceeds a predefined threshold, then the process terminates; otherwise a new chromosome would be formed again.

Xpath-DOM language definition in BNF: Although a lot of research used the Xpath technique to extract the content of Web pages, but due to the complexity of the Xpath grammar, it is difficult to perform the processing automatically. Hence, a more simple language should be important. Moreover, there exists a "last mile" problem the when the Xpath expression locates the probable position the final content could not still be distilled. So in this case, the Dom tree structure can be used to simplify the whole Xpath expression. Therefore, in this paper, we propose a simple language integrate the Xpath and Dom together called Xpath-DOM Language.

Firstly, we should review the standards of grammar specification, i.e., Backus Naur Form (BNF).

```
<XPath-DOM> ::= <XPath> "&" <DOM>
<XPath> ::= <XPath-Expr> | <XPath-Expr> <Path-Op> <XPath-Expr>
<XPath-Expr> :: = <PathLocator> <Element> | <PathLocator> <Element> "[" "@" "class=" "'" <Attrib> "'" "]"
<PathLocator> ::="/"|"//"
<Element>::=ElementTag
<Attrib>::= AttribTag
<DOM> ::= <DOM-Expr>
<DOM-Expr> ::= <DOM-Expr><op><DOM-Expr> | <DOM-Terminal> | <VarNode> "(" <Number> ")"
<op> ::="."
< DOM-Terminal> ::= "text" | "count" | "sum" | "value"
<VarNode> ::= "E" | "A" | "T"
<Number> ::= [0-9]+
```

Fig. 2: BNF grammar of Xpath-DOM language

```
▼<li sku="541877" onclick="log(7,3,541877)">
 ▼<div class="p-img">
  ▼<a target="_blank" href="http://www.360buy.com/product/541877.html">
     <img onerror="this.src='http://www.360buy.com/images/none/none_150.gif'" width="160" height="160" alt="苹果(Apple) MacBook Pro MD313CH/A 13.3寸
     4af3-af5b-a5faee429537.jpg">
    </a>
    <div class="pi8"></div>
   </div>
 ▼<div class="p-name">
  ▼<a target="_blank" href="http://www.360buy.com/product/541877.html">
     "苹果(Apple) MacBook Pro MD313CH/A 13.3寸宽屏笔记本"
     <font style="color:#ff0000" class="adwords" name="541877"></font>
    </a>
   </div>
 ▼<div class="p-price">
    <img onerror="this.src='http://www.360buy.com/images/no2.gif'" src="http://price.360buyimg.com/gp541877,1.png">
   </div>
 ▼<div class="extra">
  ▼<span class="evaluate">
     <a target="_blank" href="http://club.360buy.com/review/541877-1-1.html">已有371人评价</a>
    </span>
    <span class="reputation">(98%好评)</span>
    <span id="p541877"></span>
   </div>
 ▼<div class="btns">
    <a onclick="log(1,7,541877)" href="http://jd2008.360buy.com/purchase/InitCart.aspx?pid=541877&pcount=1&ptype=1" target="_blank" class="btn-buy">
    <input id="coll541877" type="button" class="btn-coll" value="关注" onclick="feed_publish_collect(2,541877);">
    <input type="button" class="btn-comp" value="对比" onclick="addToCompare(this,541877,'苹果(Apple) MacBook Pro MD313CH/A 13.3寸宽屏笔记本')">
   </div>
</li>
```

Fig. 3: HTML source of the example page

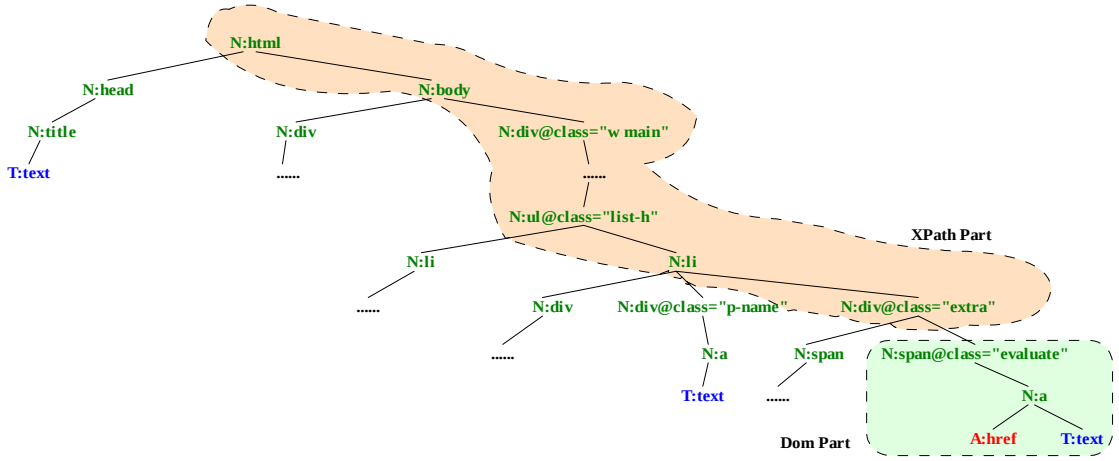


Fig. 4: DOM tree with Xpath and DOM parts

Definition 1: Backus Naur Form (BNF): BNF is a notation for expressing a language grammar as Production Rules (PRs). BNF grammar consists of the tuple <T, N, P, S> where, T is terminal set; N is non-terminals set; P is PRs set; S is start symbol.

Definition 2: Xpath-DOM: The Xpath-DOM Language whose PRs can be defined by a Context Free Grammar (CFG) in BNF as shown in Fig. 2.

Now we investigate the example shown in Fig. 1 and we can examine the html source with the help of browser as in Fig. 3 and also get the Dom tree in the form of Fig. 4. In it, green, blue and red node denotes element, text and attributes respectively. According to the above grammatical production rules as shown in Fig. 2, we could also write a two-parts expression of Xpath-DOM, the XPath part is expressed in a orange region and the DOM part in a green one. For instance, if someone wants to get the entrance URL of reviews, a expression of "/HTML/BODY/DIV [@ class = 'wmain'] //UL [@ class = 'list-h'] /LI/DIV [class = 'extra'] & N (0). N (0). A (ref). value" or simply "//UL [@ class = 'list-h'] //DIV [class ='extra'] & N (0). N (0). A(ref). value" should be provided.

Chromosome foundation and transformation: The proposed algorithm uses fixed-length chromosomes rather than variable-length. This restriction limits the

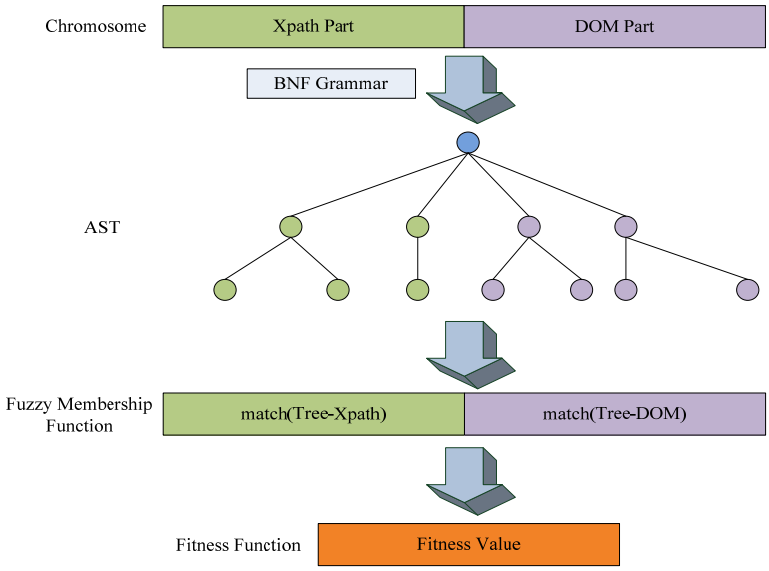


Fig. 5: Process of fitness value evaluation

creation of very large expressions decreasing also the search space. Formally, a chromosome C can be represented as binary and consists of a set of genes:

$$x = (c_1, c_2, c_l), c_l \in \{1, 2, \ldots, 255\} \quad (2)$$

where, l is the fixed length of the chromosome and the upper bound 255 limits the set number of PRs for each terminal symbol cannot exceed it.

The genetic operators of crossover and mutation are applied to the genetic population forming the next generation of chromosomes.

- **Crossover operation:** In the crossover procedure, a number of new chromosomes are created replaced by the new population with the lowest fitness value in the current generation. Usually, the crossover probability in the present implementation is set to 0.95.
  Pair of chromosomes, randomly selected parents from the current pool, is segmented at a randomly chosen point and the right-hand sub-chromosomes are exchanged. The parents are selected through tournament selection method, i.e., first a group of K>2 randomly chosen chromosomes is formed; then the individual with the best fitness in the group is selected; finally, the others are discarded.
- **Mutation operation:** In this step, a random number in an interval (0, 1) is chosen for each unit in a chromosome and each chromosome can be changed in a range (0, 255). If this number is less than or equal to the mutation rate, then the related unit is changed randomly; otherwise it is remained intact.

Fitness value evaluation: The chromosome is split into two parts: Xpath part and Dom part, which are used to construct respective features by the mapping processes as shown in Fig. 5. According to these two features, the fitness function can be defined as follows:

$$\begin{aligned} & fitness(tree_{XPathree}, node_{Dom}) \\ & = match(tree_{XPath}) \times match(node_{Dom}) \\ & = (match(text_{XPath}) \times (1 - Redundance_{XPath})) \times \\ & (match(text_{Dom}) \times (1 - Redundance_{Dom})) \\ & = \prod_{i=XPath}^{Dom} match(text_i) \times (1 - Redundance_i) \end{aligned} \quad (3)$$

where, the match() function calculate the fuzzy membership value of the parsed text in a tree node related to the object text often in a known database; and the $Redudance_{Xpath}$ computes the redundancies of the tag soup which is defined in the next section. Obviously, in the idea case, according to Eq. (3), if the Xpath part and the DOM part locate the objective tag accurately, the fitness value is 1; otherwise, if the noise information is larger than the useful very much, then this fitness could reach to 0.

EMPIRIMENTAL EVALUATION

To investigate the performance of our approach, we tested this algorithm on some real electronic commerce Website ranged from comprehensive shopping center to travel agent site. To retrieve the content pages, we implemented a theme crawler for these sites in Java, which is characterized by the future model, thread pool and blocking queue promised the concurrency and efficiency. Next, we converted the fetched pages into a XML file by the html parser of

Table 1: Xpath-DOM expression for fields of content of parsed pages

	Content					
Web site	Title	District	Address	Phone	Recommend	Comment no
www.aibang.com	//H4&N(0).N(1).text	//TD[@class='td_city'] &N(0).text	//TR&N(9).N(3).N(0).text	//SPAN[@class='img_validate']/SPAN&N(0).text	//TR&N(14).N(3).N(1).text	//DIV[@class='shoptalkcontent']&N(0).count
www.dianping.com	//DIV[@class='title']&N(0).N(0).N(0).N(0).text	//META[@name='keywords']&N(0).A(2).value	//DIV[@class='address']&N(0).N(0).N(1).text	//DIV[@class='address']&N (0).N (1).N(1).text	//DL[@class='intro']//DD/A&N(0).text	//UL[@class='remarkCate'] /LI/A/SPAN& N(0).text
beijing.fantong.com	//H1&N(0).text	-	//DIV[@id='ctinfo_Content']&N(0).N(1).N(1).text	-	//DIV[@class='tuijculined']//A[@href]&N(0).text	//DIV[@class='num1']/SPAN[@class='fred'] &N(0).text
www.gudumami.cn	//H1[@align='center'] &N(0).text	//A[@class='position_a'] &N(0).text	/BIG&N(0).N(2).text	-	-	-
www.yanwoo.cn	//H1&N(0).text	//OPTION[@value='all']&N(0).text	//TABLE[@cellpadding='0']&N(0).N(1).N(1).text	//TABLE[@ cellpadding='0']&N(0).N(5).N(1).text	-	-
biz.ziye114.com	//DIV[@class='title name'] &N(0).text	//DIV[@class='l']/A &N(1).text	//DIV[@class='xm']& N(1).N(2).text	//DIV[@class='tel']/DIV[@class='detail']&N(0).text	//DIV[@class='xm'] &N(7).N(2).text	-

Table 2: Recall, precision, and redundancy (%) of LPs and DPs

Category of site		Rc	Pc	Rd
Comprehensive shopping center site	LP	82.6	86.3	100
(360 buy, china-pub, etc.)	DP	81.3	83.9	99.7
Food consumption site (aibang, fantong, etc.)	LP	92.4	100	100
	DP	84.1	92.7	98.2
Travel agent site (kunxun, ctrip, etc.)	LP	94.7	100	100
	DP	87.2	95.5	97.9

cyberneko package so that we can use the method of Xpath and Dom. Avoid the deny configuration of the source Website, we also import proxy technique to fetch enough pages. For each site, we first downloaded 100 list pages and 100 detail pages; then these pages are parsed and extracted by the GE package EpochX v1.4.1 (a genetic programming software for research) (Castle and Beadle, 2007); finally, the extracted contents would be merged into a nested document in JSON/BSON format and stored in NOSQL database MongoDB.

In our experiments, we considered three criteria, i.e., recall ratio, precision ratio and redundancies ratio as shown in formulae (4-6), to investigate the performance of the proposed method:

$$Recall = \frac{N_{correct}}{N_{total}} \tag{4}$$

$$Precision = \frac{N_{correct}}{N_{correct} + N_{missing} + N_{mistake}} \tag{5}$$

$$Redundance = \frac{N_{useless}^{tag}}{N_{correct}^{tag} + N_{useless}^{tag}} \tag{6}$$

where,

$N_{correct}$: Number of the entities recognized correctly
$N_{missing}$: Number of the entities unidentified
$N_{mistake}$: Number of the entities recognized incorrectly
$N_{useless}^{tag}$: Number of the useless tag in the entity
$N_{correct}^{tag}$: Number of the tags relevant directly

Table 1 describes the Web sites, the content fields should be extracted and the concrete Xpath-DOM language expression generated by the GE automatically. Take the phone number of "www.aibang.com" for instance, the expression "//SPAN [@class = 'img-validate'] /SPAN&N (0).text" denotes that in absolute path "/SPAN" under the relative path "//SPAN [@class = 'img-validate']" there existing a child node N (0) and its text is our object. The "-" expresses the failure of extraction may due to the corresponding structure never exists or could not generate the proper expression of that objective. In order to show convenient, we use a form of "[@class = 'some attribute'] "to substitute a meaningless internal expression of GE.

Table 2 summarizes the performances of all experiments ranged from comprehensive shopping centers to travel agent sites. It is clear from Table 2 that this approach has high precision and redundancy, also the recall is slightly low. As it not need any given template beforehand, the high precision and redundancy means the result could be obtained without artificial intervention and reconstruction. In other words, this method is considerable automatic and intelligent. Compared to the approaches only use Xpath, the hybrid of Xpath and the Dom shorts the length of the expression and reduces the solution space greatly. Moreover, the results also show that the proposed algorithm not rely seriously on the concrete Dom tree structure, although it differs the list pages and content pages slightly. So it could be applied in broad scope of pages.

CONCLUSION

In this srudy, we proposed a new approach to extract structured data from electronic commerce Website pages. Although the problem has been studied

by several researchers for a long time, existing techniques are either inaccurate or not automatic. Our method does not only make any assumption about the structure of the pages, but also needs any temple difficult to provide. A reduced language integrating Xpath and DOM techniques is given to generate the solution of parse in a BNF grammar form, which is used in the Grammatical Evolution (GE) approach. Empirical results on several real Web pages show that the new proposed technique can segment data records and extract data from them accurately, automatically and flexibly.

ACKNOWLEDGMENT

This study is supported by the Fund of Guangxi Autonomous Region Education Department General Project (201204LX409), Guangxi Autonomous Region Education Department General Project (200103YB137).

REFERENCES

Brabazon, A. and M. O'Neill, 2003. A Grammar Model for Foreign-Exchange Trading. In: Arabnia, H.R., *et al*. (Ed.), Proceeding of Internaternational Conference on Artificial Intelligence. CSREA Press, 2: 492-498.

Castle, T. and L. Beadle, 2007. Epochx: Genetic Programming Software for Research. Retrieved from: http://www.epochx.org.

Chang, C.H., C.N. Hsu and S.C. Lui, 2003. Automatic information extraction from semi-structured web pages by pattern discovery. Decis. Support Syst., 35(1): 129-147.

Collins, J.J. and C. Ryan, 2000. Automatic generation of robot behaviors using grammatical evolution. Proceeding of AROB, 5th International Symposium on Artificial Life and Robotics, pp: 351-354.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge, Mass.

McKeown, K., R. Barzilay, J. Chen, D. Elson, D. Evans, J. Klavans, A. Nenkova, B. Schiffman and S. Sigelman, 2003. Columbia's newsblaster: New features and future directions. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4. Association for Computational Linguistics Morristown, NJ, USA, pp: 15-16.

O'Neill, M. and C. Ryan, 2001. Grammatical evolution. IEEE T. Evolut. Comput., 5: 349-358.

O'Neill, M. and C. Ryan, 2003. Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language of Genetic Programming. Kluwer Academic Publishers, Vol. 4.

Qiu, T.F. and T.Q. Yang, 2010. Automatic information extraction from E-commerce web sites. International Conference on E-Business and E-Government (ICEE), China, pp: 1399-1402.

Reis, D.C., P.B. Golgher, A.S. Silva and A.F. Laender, 2004. Automatic web news extraction using tree edit distance. WWW'04: Proceedings of the 13th International Conference on World Wide Web. New York, USA: ACM, pp: 502-511.

Ryan, C., M. O'Neill and J.J. Collins, 1998. Grammatical evolution: Solving trigonometric identities. Proceeding of Mendel, 4th International Mendel Conference on Genetic Algorithms, Optimization Problems, Fuzzy Logic, Neural Networks, Rough Sets, pp: 111-119.

Zhai, Y.H. and B. Liu, 2007. Extracting web data using instance-based learning. WWW, 10(2): 113-132.

Ziegler, C.N. and M. Skubacz, 2007. Content extraction from news pages using particle swarm optimization on linguistic and structural features. WI'07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, pp: 242-249.