Research for Information Extraction Based on Wrapper Model Algorithm

Xu zhiwei
Department of Computer Science
Chang Chun University
Chang Chun, china
xuzhiwei2000@sina.com

Wang xinghua
Department of Information Technology
BANK OF JILIN CO., LTD
Chang Chun, china
wangxinghua@sina.com

Abstract—**Mainly on data-intensive Web site research experiment. In the web pages of the automatically generated wrapper method of research-based information extraction, the main job is to make the page tree matching algorithm, the sample tree and the tree wrapper DOM tree matching two pages compared to the first to discover the page selection mode, producing the primary template, and then self-correction of primary template found iterative model, and finally generate the page wrapper method. The wrapper generation process does not require human intervention to achieve a fully automated completion. Experiment with satisfactory results.**

Keywords- information extraction; wrapper; DOM tree; match technology

I. INTRODUCTION

With the popularization of computer technology and network technology development, Web has been developed into a huge storehouse of information has become increasingly important and the most potential of global information transfer and sharing of resources. However, you want to quickly and accurately from the vast amounts of resources to find the needed information and is applied in other programs, it has become a major challenge. Therefore, the need for the application of information extraction technology from a large number of semi-structured information extracted structured in line with the theme of data. As the HTML page is mainly for browsing, not for the manipulation and use for which the data is difficult to directly use the application. Therefore, the data extracted from web pages and passes them to the application, is still a complex, difficult but interesting task.

II. INFORMATION EXTRACTION TECHNOLOGY OVERVIEW

A. The concept of information extraction

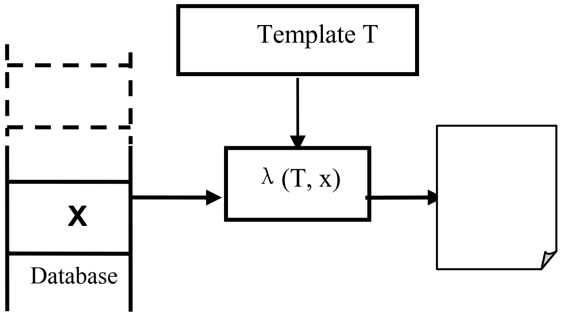
Information extraction (IE) refers to the select information from the text and its structure-based (such as the relationship between the form of a database) that process. (Grishman 1997)

Wrapper is an ability to extract data from HTML pages out, and will restore them to structured data, software programs.

Basic Type denoted by B, said a token. Token is the basic unit of text, definition of token as a text string or HTML tag.

B. Create Template of Page

Examples of the type and type are template. X for the database set up a data object, T is a page template, λ for a coding function, it will use the template T data object x embedded in an HTML page. T using a template encoded on the x after the page recorded as λ (T, x). Page creation process is to use coding function λ of the data in the database object x encode and output process (FIG 2-1). λ (T, x) is an example of T. the structure of the data x coding to HTML documents, the data structure of the information would be completely lost. Wrapper is the process carried out by the inverse process will be the page in the semi-structured information into a re-structured form.



C. Wrapper Generation

Figure 1. Create Template of Page

Wrapper generation and data extraction can be formalized is defined as: given a n-sample of the same type of a collection of HTML pages P, pi ∈ P, page pi = λ (T, xi) (1 ≤ i ≤ n), that is, P by unknown template T and the data sets (x1, ..., xn) generated. Wrapper generation and data extraction from the page collection of P derived template T and the data sets (x1, ..., xn) process.

978-0-7695-4043-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCRD.2010.141

III. PAGE LEXICAL ANALYSIS

Said that the theme for the Web page contents of a string of data are included in a pair of HTML tags is the information extraction process to extract data. We use a PCDATA text symbols to represent these strings.

Definition 3.1 Σ is a page set up an open-label symbols of all elements in a collection $\overline{\Sigma}$ of all elements of the page closing tag symbols set, Σ = (</ a> | <a> ∈ Σ), pages alphabet is defined as: ΣU $\overline{\Sigma}$ (PCDATA)

Definition 3.2 The structure of well-defined sequence of symbols: Let SL is a sequence of symbols on the page the alphabet, if the SL line with context-free grammar G, claimed that SL is a well-structured sequence of symbols.

G=({S,X,XD}, ΣU$\overline{\Sigma}$U{PCDATA}，R，S)，One rule set R =：

$$S \to aX_D\overline{a} / XX_D X$$

$$S \to a\overline{a} / aX\overline{a} / XX$$

$$S \to aX_D\overline{a} / XX_D / X_D X / PCDATA$$

a is an open-label symbol in Σ, $\overline{a}$ is $\overline{\Sigma}$ closed with a corresponding label symbols.

HTML pages can be viewed from the pages of the label alphabet symbols and strings composed of symbols stream. Page lexical analysis process will parse the page's DOM tree in the grounds of the page the symbols of the alphabet well-structured sequence of symbols, that is, the sequence of each open-label symbols are corresponding to the closing tag symbols, and the opening and closing label symbol with the right level of nesting.

Page lexical analyzer from the DOM root node, the depth of its priorities and begin to traverse them. Traversing down the element node in the element is generated when the open-label symbols, traverse return the element's closing tag is generated symbols; traverse to the node then generate a text string symbol PCDATA. Traverse ignoring <font> <b> elements such as text-decoration. After traversing the sequence of tokens generated pages.

IV. PAGE TREE MATCHING ALGORITHM

In the data-intensive Web sites is usually generated automatically: the data is stored in the back-end database management systems, HTML pages use a script to make generation (IE program-based database content). In this article, research structure and data extraction process: "In view of this group the sample HTML pages belonging to the same level, to find the nested type of the source database and web pages extracted from the source data, the study found the page selection mode and iterative model."

Set up by the same one page template generated k-(k ≥ 2) a sample page example of the HTML pages expanded to two kinds of nodes to represent the iteration and choice:

Iterator nodes: (S) + corresponds to an Iterator for the root, to S as a sub-tree tree;

Option Node: (S)? corresponds to an Option for the root, to S as a sub-tree tree.

By matching to find the page is optional. The main processing of the matching process is relatively the same time, the input and amendment of the difference between the two pages tree node, and generate a minimum page tree. Match two pages tree, which is called a wrapper tree, denoted by Tw, the other is called the sample tree, denoted by Ts.

A. The page tree matching algorithm implementation process

The algorithm is applied matching technology, through the token sequence of two HTML pages (wrapper and sample) to match, either take one of the wrapper as the original sample (TW), and then continued with the sample (TS) to match the comparison between solution mismatch to discover common regular expressions. If the syntactic analysis process by addressing all of the does not match the automatically generated wrapper, then the algorithm is successful.

In the syntactic analysis does not match the definition of the two kinds: a string does not match (String mismatches), a label does not match (Tag mismatches) which does not match the label again in two points, namely: selection does not match the (Option mismatches) (may not exist and can exist) and the iteration does not match (iterate mismatches).

1) Primary Template Generator.

In this process, the first page and sample pages wrapper separately standardized treatment, the purpose is to make HTML pages take place independently of each label for easy comparison match.

a) Normalized processing algorithm steps described in:

The establishment of a temporary file (in order to import a template, the same two pages);

Read into the HTML page line by line

If you read the HTML code "" "or" "" when, on the importation of a newline character;

Otherwise, write directly;

Changes to the temporary file, generate standardized documents, the two consecutive line breaks to delete a row.

b) Solution does not match the generated primary template:

The string does not match the definition:

Matching the discovery process in a specific address does not match the two types do not match, namely the string does not match and labels do not match, how to distinguish the string does not match or the label does not match the problem is very easy to solve, in this paper, we The two pages into standardized treatment for progressive matching, because all the labels are a common format, such as <a> </ a>, therefore, in the match, if not match appears in the "<"node of the line, we have defined as a label does not match, in addition are defined as the strings do not match. In the generated template, as long as the string has been established that the node does not match the line, you use # PCDATA string instead of the Bank.

Tag does not match the select does not match the definition:

The string does not match the definition has been described how to distinguish the string does not match or the label does not match, so the matching process in a wrapper the first line of the first page with the sample does not match a row took place, and this article contains a line or a "< "node of the line, so we determine labels do not match. For determining the label does not match the line of nodes, how to distinguish between the choice does not match or does not match the iteration is a problem, this study focused on this issue based on the assumed long as it is found that the label does not match the first and foremost as a is to choose does not match the processing, the application method of cross matching cross-comparison.

Criss-cross method: the assumption of selection for the two pages do not match the line number is "a". Separate wrapper page first page a line with the sample the first line of matching a +1; sample page first page a line with the wrapper the first line of matching a +1 to see if a successful match. If it is a wrapper page is the first line of the page with the sample matched the success of the first line of a +1, put a sample page of the first line of the node is defined as to select a node, then using symbols (S)? Place in the node part of the emergence of choice. Similarly vice versa.

2) Ultimate Template Generator.

Has been described in section 4.1.1, primary template will generate the iteration does not match the label does not match the definition also choose not to match, an error in this section we will discuss how to solve this problem.

Iterative definition:

In the HTML page analysis, such as a node, there are several parallel sub-nodes, the nodes of the existence of such structures is defined as the iteration node. For example, for the same detailed information described in this book can also exist a number of different information, such as there are a number of different versions of information. This information appears in a book within the larger structure of the external iteration, there has been little internal iteration, the definition of the existence of such nodes in the structure of the internal iteration.

Validation in the primary template found there will be some mistake does not match the external and internal non-matching problems. Based on the HTML tag language features we can see there is an open-label <*>, there must be a necessary label </ *> corresponding. Us to resolve the error template is based on an idea first does not match the deal with internal iteration, generate a sub-iteration template; re-treatment does not match the external iteration to generate the ultimate template is to deal with all the non-match (choose not to match and iteration does not match) the template after the final wrapper.

Iteration does not match the internal processing:

Generated primary template (# PCDATA)? Find, compare (# PCDATA)? The upper and lower labels for the above is </ *> the following tag is <*>. If not, would like to go downward. If it is, to his post<*> label down to find the nearest corresponding </ *>label, the <*> ... ... </ *> the string between the deleted in its entirety, and to (# PCDATA)? Replace (# PCDATA) +, the end.

External iteration does not match the processing:

</*><*> Find the template in line with a pair of label, after finding the back of the label <*> to find down to the nearest corresponding end tag </*>, the <*>... ... </ * >the string between all deleted, and then up to find<* >... ...< / * >and to< * >... ...< / * >with the means of expressing iteration, namely, (<* >... ...</ *>)+, end.

V. WRAPPER GENERATION

Samples of the wrapper tree and turn the page to match the page tree compared to a sample page up with all the matches were finished. Finally the resulting wrapper tree is that we can repeat the pattern expressed in the sample page and choose the smallest model of the page tree. The page tree output corresponding regular expression, that is the final regular expression wrapper.

VI. IMPLEMENTATION AND EXPERIMENTS

To validate the algorithm match described above, we have developed a prototype of the wrapper generation system and used it to run a number of experiments on HTML sites. The system has been completely written in Java.

In order to make a comparison, we downloaded from RISE, a repository of information sources from data extraction projects

	Recall ratio	**Precision ratio**	**F**
Essential_Information(EI)	97.14%	98.55%	97.87%
Educatioanl_Backgroud(EB)	96.45%	97.01%	93.24%
Contact(C)	73.91%	91.07%	81.60%
Foreign_Language(FL)	95.38%	98.41%	96.87%
Job_Intentions(JI)	60.87%	84.00%	70.59%
Treatment_Intentions(TI)	61.54%	83.05%	70.33%

TABLE I. INFORMATION EXTRACTION RESULTS

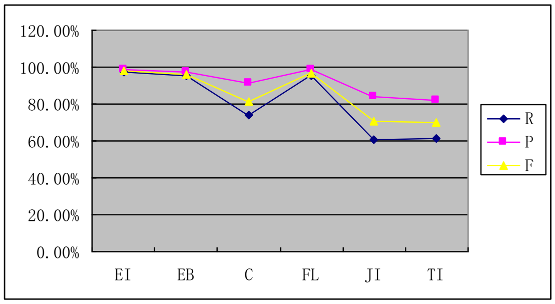


Figure 2. Experimental setup Comparison Chart

The results of information extraction can be seen that basic information, educational background, foreign language situation in the recall rate, accuracy rate, and F mean are more ideal, while the job search intention and the intention of the extraction treatment of a lower recall rate. Was mainly due to EI, EB, FL information contained in the data is rich in description of the types of information is relatively clear and obvious effect.

REFERENCES

[1] M Craven, D. DiPasquo, D. Frei tag et al. Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence, 2000, 118(1-2):69-113.

[2] R Gaizauskas, Y Wilks, Information Extraction: Beyond Document Retrieval. Journal of Documentation, 1997.

[3] K Lerman, Minton S N, C A Knoblock. Wrapper Maintenance:A Machine Learning Approach. Journal of Artificial Intelligence Research, 2003, 18:149-181.

[4] C VALTER, G IANSALVATOREM ECCA. Road Runner: Towards Automatic Data Extraction from Large Web Sites. In Proceedings of the 27th International Conference on Very Large Database, Roma. Italy, 2001.

[5] S Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning, 1999, vol. 34(1): 233-272.

[6] Sun tie li, Li zhiying, Algorithm Research for the Noise of Information Extraction Based Vision and DOM Tree 2009 International Symposium on Intelligent Ubiquitous Computing and Education （IUCE 2009）

[7] Sun tie li, Li zhiying, Research and Evolution for Information Extraction Based on Wrapper Journal of Northeast Normal University (Natural Science Edition) 2008.6