A CRF-based Approach for Web Object Extraction

Rui Liu¹, Rui Xiong², Kun Gao³
State Key Lab of Software Development Environment, Beihang University,
No.37 Xueyuan Road, Haidian District, Beijing, 100191, P.R.China
{l1u1rui, xiongrui, gaokun}@nlsde.buaa.edu.cn

Abstract—A method for extracting Web object is presented in this paper. Firstly, Web object blocks are obtained by blocking the web page and calculating the information entropy of it. Then it uses Conditional Random Field model as a probability and statistics model, and builds a series of feature templates according to the characteristics of objects themselves. Feature functions are generated based on the result of Chinese word segmentation and feature templates. It uses a limited memory BFGS algorithm to estimate parameters of the model, and labels property sequences of Web object blocks by Viterbi algorithm. Experiment result shows that the proposed method is an effective way to extract science data.

Keywords- Information Extraction, Web Object, Machine Learning, Conditional Random Field

I. INTRODUCTION

Web object [1] extraction has become an important and valuable research field in recent years. As the basic data unit of web information collecting, indexing and ranking, Web object corresponds to the thing that can be recognized in real world, such as cameras, and other products on the Web pages. Every type of objects has their own properties that identify and describe themselves. As shown in Fig. 1. There are two books in a list. From the object point of view, they are both objects with their own properties such as name, author and price. Web object extraction can extract user interested information from Web pages and express them in a specific structure, such as table in database. So, people can analyze these data based on the powerful database technology. Comparing to general search engine, it can provide more accurate, targeted information, and facilitate the following objects ranking and association mining.

Web object extraction has to solve two problems: recognize the web objects in Web pages, and extract the information about it. Approaches based on Nature Language Processing (NLP), such as WHISK [2], generate extraction rules by parsing, semantic annotation, recognition of proper noun and other steps. These methods don’t think about the structure of the Web page. They are good at deal with the free text, such as a notice of seminar, in Web pages. Some approaches use wrappers, such as STALKER [3]. A wrapper is an application used to extract data in Web pages. They depend on the format of data not the linguistic feature of data. Usually, a wrapper is specific to a data source. Different types of Web pages need different wrappers. Approaches based on Html structure of Web page, such as [4], find out the pattern corresponds to a Web object by observing the Html syntax tree, layout style and so on, and generate extraction rules based on these patterns. Some researches, such as [5], construct the ontology library by experts for specific domain, and generate the extraction rule base on it. But it is costly to build and maintain ontology library. Some other methods are based on statistical machine learning, such as [6]. These methods can learn data feature through training. They are independent from the site structure, have good adaptability and high precision.

In this paper, we proposed a Web object extraction method based on CRF. It can automatically extract Web objects from a variety of heterogeneous websites to provide value-added services, such as providing meta-search. The method has two steps. First, we find interesting Web objects with entropy pruning algorithm to avoid noise data. Second, we employ an object property extraction algorithm based on CRF to find the correct property information from object blocks obtained in the first step. We used it in our project - National Science and Technology Portal of China and got a satisfied result.

II. WEB OBJECT BLOCK MINING

Web objects are often found in table or list on Web pages in a variety of forms. Statistics show that about 52% of the pages contain a table or a list of products [1]. On the other side, Web pages contain not only the Web objects, but also some noise regions such as advertisements, search panels and navigator bars, which will decrease the precision of extraction. Web Object Block Mining will detect the semantic blocks in the Web pages, and calculate the entropy of each block; then, prune the noise regions to find the Web objects in the Web page. It contains the following steps (Fig. 2):

Fig. 1. Two Web objects in a HTML list

978-1-4244-5539-3/10/$26.00 ©2010 IEEE
III. WEB OBJECT PROPERTY EXTRACTION

We have got the Web object blocks through Web Object Block Mining. The one of the blocks we extract may be as follows:

"Scanning Electron Microscope&&-NL30&&Shanghai-Shanghai Material Research Institute&&XL30&&291&& http…”

We look forward to the following form of mapping results:

"Scanning Electron Microscope/N&&-
NL30/O&&Shanghai-Shanghai/O&&Shanghai Material Research Institute/P&&XL30/T&&291&& &http…/L”


Now, the problem we faced is: Given object property sequence $E = \langle e_1, e_2, ..., e_r \rangle$, find optimal label sequence $L = \langle l_1, l_2, ..., l_r \rangle$, where $l_i = \langle a_1, a_2, ..., a_8 \rangle$. Thus, Web object property extraction is transformed into a sequence labeling problem.

As shown in Fig. 3, our method contains two stages: training and labeling. In training stage: First, Web object blocks are pre-processed to form formatted sequence which is going to be labeled. Next, Chinese word segmentation is used. Then, a CRF model is established according to the feature functions extracted out based on feature templates. Finally the parameters of the feature functions in the model are estimated. Through these steps, the Web object property information extraction models for particular fields are
achieved. We can use them to extract useful information in labeling stage.

In the following sections, we will use large scientific instruments and equipments as an example to describe the key steps of our method.

A. Design of Feature Template

One of the main advantages of CRF is that it can bring a large number of different features into the model with no need to comply the complete independence assumption. Template is the consideration of specific location and specific information of the context. Template function is used to find out the differences between property fields, that is, determine which property the words belongs to. The selection of feature templates is based on this principle. Different objects have different properties. Feature selection must define appropriate feature set for the model according to the specific extraction.

In the specific task of large scientific instruments and equipments properties extraction, seven properties have to be extracted: name, type, origin, supplier, price, purchase time and detailed information link. According to these properties, four feature templates will be built: atomic feature template, external dictionary feature template, state transition feature template and other state feature template.

1) Atomic Feature Template

Atomic feature is the spelling or character form of the field itself in the input sequence. Table I shows the atomic feature template (the examples are translated into English for a better understanding). The WordFeature template represents the features of the current word. After Chinese word segmentation, all different segmented words are stored in the WORD dictionary for the specific model and through word segmentation, all different segmented words are stored. The selection of feature templates is based on this principle. Different objects have different properties. Feature selection must define appropriate feature set for the model according to the specific extraction.

In the specific task of large scientific instruments and equipments properties extraction, seven properties have to be extracted: name, type, origin, supplier, price, purchase time and detailed information link. According to these properties, four feature templates will be built: atomic feature template, external dictionary feature template, state transition feature template and other state feature template.

2) External Dictionary Feature Template

External dictionary feature is used to describe some feature words and characters which have a record in the external dictionaries. These external dictionaries include: scientific instrument dictionary, country name dictionary, Chinese province and city name dictionary.

3) State Transition Feature Template

State transition feature is the relationship between current state and the adjacent one (or ones). In our task, we find that some property fields have strong sequential relationship. For example, “instrument type” field always follows the “instrument name” field. Through state transition feature, we can put this potential information into the model. State transition feature is defined as the feature function of CRF. For example, the function used to describe “instrument type” following “instrument name” can be defined as:

\[
f_k(y_{i-1}, y_i, x) = \begin{cases} 1 & \text{if } y_{i-1} = \text{name} \text{ and } y_i = \text{type} \\ 0 & \text{otherwise} \end{cases}
\] (3)

4) Other State Feature Template

Other state feature records the location information of the current state. For example, “instrument name” often appears at head of a sequence, whereas “URL” is often at the end of a sequence.

B. Parameter Estimation

CRF model can be mathematically expressed as:

\[
P(y|x, \Theta) = \frac{1}{Z(x)} \exp \left( \sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x) \right)
\] (4)

Where \( \Theta = \{\lambda_1, \lambda_2, ..., \mu_1, \mu_2, ...\} \) is a set of all parameters. For a given training set, a series of feature functions, that is \( f_k \) and \( g_k \) in (2), can be generated through the feature templates. Each feature function has its unknown parameters \( \lambda_k \) and \( \mu_k \).

The target of parameter estimation is to determine a specific parameter sequence \( (\lambda_1^*, \lambda_2^*, ..., \mu_1^*, \mu_2^*, ...) \) for a given training set \( D = \{x^i, y^i\} | i = 1, 2, ..., N \} \), making the probability get the max value.

We use maximum likelihood estimation method to do parameter estimation. Assume D is a set of sample data of training data, \( p(x, y) \) is the empirical probability of \( (x, y) \) in D. For the conditional model \( p(y|x, \Theta) \), the maximum likelihood function is:

\[
L(\Theta) = \sum_{i} p(x, y) \left( \sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x) \right)
\] (5)

\( L(\Theta) \) is a smooth concave surface in the whole parameter space \( \Theta \), which means that it exists a \( \Theta \) making the maximum likelihood function achieve the global optimum value and the gradients or vectors of partial derivatives of every parameters in each \( \Theta \) are zero.
We use limited-memory BFGS (Broyden Fletcher Goldfarb Shanno, L-BFGS) algorithm [9,10], one of the limited memory quasi-Newton method to implement parameter estimation.

C. Determine the Best State Sequence

When the parameter estimation is completed, the model can be used to extract the properties of similar Web objects. It will be done by labeling the given input sequence properly and finding the best state sequence $y^*$ through training model.

$$y^* = \arg \max_y P(y \mid x, \Theta)$$  \hspace{1cm} (6)

In the formula, $\Theta = \{\lambda, \alpha, \ldots, \mu, \nu, \ldots\}$ is the trained model containing all the parameters. To get it, one basic way is traversing all possible state sequences and inputting them into Formula (2) to get the probability $P$. Then, the result of $y^*$ can be got. But this method has its deficiency. Given the number of observation sequences is $N$ and the number of states is $M$, then the time complexity is $O(M^N)$. It is clearly unfeasible when $N$ and $M$ are very large.

We adopt an efficient algorithm, Viterbi algorithm [11], to calculate the state sequence. Viterbi is a dynamic programming method similar to forward-backward algorithm in HMM. It calculates the global optimal state sequence by recording every optimal state sequence in every moment, so it can avoid repetitive calculations in traversal process.

IV. EXPERIMENTS


To examine the practicability of CRF-based Web object properties extraction method, this paper chooses Maximum Entropy Model which is commonly used in Web information extraction for comparison. The experiments were carried out under the same environment. The times of iterations are 50. Extraction results are shown in Table II and Table III. The first column is the name of each field of web object property. The second column is the number of fields labeled correctly. The third column is the total number of fields labeled by this method. The fourth column is the total number of fields in our test set. And last three columns represent the precision, recall and F-measure.

<table>
<thead>
<tr>
<th>Field</th>
<th>correct labeled sum</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>1347</td>
<td>1460</td>
<td>1428</td>
<td>92.26</td>
</tr>
<tr>
<td>type</td>
<td>707</td>
<td>943</td>
<td>822</td>
<td>74.97</td>
</tr>
<tr>
<td>origin</td>
<td>471</td>
<td>477</td>
<td>753</td>
<td>98.74</td>
</tr>
<tr>
<td>supplier</td>
<td>1410</td>
<td>1680</td>
<td>1428</td>
<td>83.92</td>
</tr>
<tr>
<td>price</td>
<td>745</td>
<td>746</td>
<td>753</td>
<td>99.86</td>
</tr>
<tr>
<td>time</td>
<td>606</td>
<td>606</td>
<td>606</td>
<td>100.00</td>
</tr>
<tr>
<td>link</td>
<td>1428</td>
<td>1428</td>
<td>1428</td>
<td>100.00</td>
</tr>
</tbody>
</table>

From the tables, we can see that the precision of CRF model is much higher than MEM on “name” field. That is because MEM models do not take into account the feature of relationship between the fields in sequence. But on the “type” field, CRF is about 20% lower than MEM. The reason is that the feature of the “type” itself is the crucial factor. When our method considers other features, it introduces more noises. However, on all fields, CRF performs better. Compared to precision, on the “name", "type" "origin" and "supplier" fields, recall of CRF models are even higher.

Fig. 4 is the F-measure comparison chart of CRF Models and Maximum Entropy Models. It combines precision and recall of the model labeling, reflecting the overall performance of the large-scale scientific instruments and equipments extraction.

![Fig. 4. F-measures comparison between CRF and MEM](image-url)
an important role in the large-scale instruments and equipments information extraction.

V. CONCLUSIONS

In this paper, we propose a high-performance method for automatic Web object extraction. It finds relevant topic region with entropy pruning algorithm, and uses Conditional Random Fields model as the probability and statistical model for property extraction. It can freely choose the features and avoid the strict independence assumption and labeling bias problem. Experimental results show that it can achieve better performance than MEM method. In the future, we will try to combine visual information with our method, and work on extracting information from Web pages with many JavaScript codes.

ACKNOWLEDGEMENTS

The research is supported by the fund of the State Key Laboratory of Software Development Environment SKLSDE-2009ZX-12.

REFERENCES


