

Extracting Objects from the Web

Zaiqing Nie¹ Fei Wu² * Ji-Rong Wen¹ Wei-Ying Ma¹
¹Microsoft Research Asia ² Tsinghua University
{znie,jrwen,wyma}@microsoft.com wufei98@mails.tsinghua.edu.cn

Abstract

Extracting and integrating object information from the Web is of great significance for Web data management. The existing Web information extraction techniques cannot provide satisfactory solution to the Web object extraction task since objects of the same type are distributed in diverse Web sources, whose structures are highly heterogeneous. In this paper, we propose a novel approach called Object-Level Information Extraction (OLIE) to extract Web objects. This approach extends a classic information extraction algorithm, Conditional Random Fields (CRF), by adding Web-specific information. The experimental results show OLIE can significantly improve the Web object extraction accuracy.

1 Introduction

This paper studies how to automatically extract object information from the Web. The main challenge is that objects of the same type are distributed in diverse Web sources, whose structures are highly heterogeneous. For instance, information about “paper” objects can be found in homepages, PDF files, and even online databases.

Although it is possible to combine existing Web information extraction techniques to construct a toolkit to extract object from some template-generated Web pages. We think this is not a practical solution, since attribute values of an object are extracted from various Web sources independently, it is required to learn a template for each Website.

Another tightly related work is classic information extraction from plain text document [1]. However, these methods are originally designed for processing plain texts and not for Web pages, and thus cannot be directly applied to the Web object extraction task. Of course, we can transform each Web page into a plain

text document by removing HTML tags and other irrelevant codes. But treating Web pages as plain text documents is unwise since some important Web-specific information for object extraction, such as page structure and layout, is lost.

The advantage of classic IE algorithms is their capability of handling heterogeneous data sources and integrating information extraction and object identification in a uniform framework, while Web IE takes advantage of the Web-specific information, e.g. tags and layouts, to extract objects. In this paper, we present an *object-level information extraction (OLIE)* approach which can effectively extract Web objects from multiple heterogeneous Web data sources. Our basic idea is to extend a classic IE algorithm, *Conditional Random Fields (CRF)*, by adding Web-specific features. So our method is essentially a combination of Web IE and classic IE. More specifically, besides text, we found that there are other two kinds of Web information, namely visual information on the Web pages and structured information from Web databases, are of particular importance for Web object extraction.

2 Problem Formulation

The Web object extraction problem is motivated by Libra, a scientific literature search engine that we are developing[3].

2.1 Object Blocks and Elements

Web Objects & Attributes: We define the concept of *Web Objects* as the principle data units about which Web information is to be collected, indexed and ranked. Web objects are usually recognizable concepts, such as authors, papers, conferences, or journals which have relevance to the application domain. Different types of objects are used to represent the information for different concepts. We assume the same type of objects follows a common relational schema: $R(a_1, a_2, \dots, a_m)$. Attributes, $A = \{a_1, a_2, \dots, a_m\}$, are properties which describe the objects, and key at-

*This work is done when the author is visiting Microsoft Research Asia.



Figure 1. Four Object Blocks Located in a Web Page and Five Elements Shown in the Bottom Block

tributes, $A_K = \{a_{K1}, a_{K2}, \dots, a_{KK}\} \subseteq A$, are properties which can uniquely identify an object.

Object Blocks & Elements: The information about an object on a Web page is usually grouped together as a block, since Web page creators are always trying to display semantically related information together. We define the concept of an *object block* as a collection of information within a Web page that relates to a single object. Given an object block found on a Web page, we further segment it to atomic extraction entities called *object elements*. In this way, the object block E_i is converted to a sequence of elements, i.e. $E_i = \langle e_{i1}e_{i2} \dots e_{iT} \rangle$. Each element e_{ij} only belongs to a single attribute of the object, and an attribute can contain several elements. Figure 1 shows four object blocks located in a Web page generated by Froogle and five elements located in the bottom block. With the help of data record mining techniques such as [2], we can automatically detect the object blocks from a Web page.

2.2 Web Object Extraction

Given an object block $E_i = \langle e_{i1}e_{i2} \dots e_{iT} \rangle$, and its relevant object schema $R(a_1, a_2, \dots, a_m)$, we need to assign an attribute name from the attribute set $A = \{a_1, a_2, \dots, a_m\}$ to each object element e_{ij} to determine the corresponding label sequence $L_i = \langle l_{i1}l_{i2} \dots l_{iT} \rangle$. If the object block E_i and a previously extracted object O_n in the database refer to the same entity, we integrate O_n and the labeled E_i together. The key attributes A_K are used to decide whether they refer to the same entity. The combined labeling and integration inference is called Web object extraction.

After locating an object block on Web pages and segmenting it to an object element set, the labeling operation can be treated as a sequence data classification problem. Please see [4] for a detailed discussion

on the sequence characteristics between the elements in an object block. To the best of our knowledge, the Conditional Random Fields(CRF) model is among the most popular and effective methods for this task [1]. So, we select the CRF as the base model and extend it for Web object extraction.

3 Object-Level Information Extraction

As stated above, our goal is to incorporate all available information to assist the Web object extraction. The basic CRF model can not meet this requirement, since it models the label sequence probability only conditioned on the element sequence $E = \langle e_1e_2 \dots e_T \rangle$, and no object identification is performed. We introduce a novel object-level information extraction approach called OLIE. Our OLIE approach uses an Enhanced CRF (ECRF) model. ECRF extends the basic CRF model by introducing two variations.

First, we modify the label sequence probability to condition on not only the element sequence, but also available databases,

$$P(L|E, D, \Theta) = \frac{1}{Z_E} \exp \left\{ \sum_{t=1}^T \sum_{k=1}^N \lambda_k f_k(l_{t-1}, l_t, E, D, t) \right\} \quad (1)$$

where, E is the object element sequence, and it contains both the text and visual information. D denote databases which store structured information. $f_k(l_{t-1}, l_t, E, D, t)$ is the new feature function based on all the three categories of information.

There are cases when we have sufficiently high confidence that some object element e_t should have certain label. For instance, the cases may be that good matches between e_t and key/important attributes of records in databases are found, or that e_t has a high enough element emission probability for some attribute. For example, if the following statistics holds, $p(l_t = \text{"conference"} | e_t \text{ contains "in proceedings of"}) = 0.99$, and current e_t is "in proceedings of SIGMOD04", it is almost definite that *conference* is the label. These constraints can be used to guide the solution searching progress to find the optimal label path correctly and quickly. This leads to our second variation for the basic CRF. Specifically, we first compute the confidence $c_t(a_i)$ that e_t belongs to certain attribute a_i based on some feature functions. If the confidence is high enough ($c_t(a_i) > \tau$), we modify the induction formula of Viterbi algorithm as follows,

$$\delta_t(l) = \begin{cases} \max_{l'} \left\{ c_t(a_i) \cdot \delta_{t-1}(l') \exp \left[\sum_{k=1}^N \lambda_k f_k(l', l, E, D, t) \right] \right\} & l = a_i \\ \max_{l'} \left\{ (1 - c_t(a_i)) \cdot \delta_{t-1}(l') \exp \left[\sum_{k=1}^N \lambda_k f_k(l', l, E, D, t) \right] \right\} & \text{others} \end{cases} \quad (2)$$

if $c_t(a_i) \leq \tau$, the induction formula is the same as (3).

Based on ECRF, our OLIE sufficiently utilizes all available information to assist the extraction for Web objects. Because object identification is performed during this process, a bidirectional communication among object blocks and records of databases is achieved, which leads to a combined information extraction and integration.

4 Experiments

The OLIE approach proposed in the paper are fully implemented and evaluated in the context of *Libra*. Two types of Web objects are defined in the experiments: papers and authors. We use instance accuracy to evaluate the performance of our OLIE approach. Instance accuracy is defined as the percentage of instances in which all words are correctly labelled.

4.0.1 Datasets

Paper Citations: We took the citation dataset derived from the Cora project for testing. It contains 500 citations and we used 300 for training and the remaining 200 for testing. 7 attributes of paper objects are extracted: Author, Title, Editor, Booktitle, Journal, Year, and Others.

Paper Headers: We randomly selected 200 papers in the Cora dataset and downloaded them from the internet. We used 100 papers for training and the remaining for testing. 9 attributes of author objects are extracted: Name, Affiliation, Address, Email, Fax, Phone, Web URL, Degree, and Others. 4 attributes of paper objects are extracted: Title, Author, Abstract, and Others.

Author homepages: We randomly collected 200 computer scientists' homepages from the internet. Compared with previous two datasets, this dataset is more general and flexible. 11 attributes of the author objects are extracted: Name, Affiliation, Designation, Address, Email, Phone, Fax, Education, Secretary, Office, and Others. We randomly selected 100 homepages for training and the remaining 100 for testing. .

ACM Digital Library: ACM Digital Library is online Web database with high quality structured data, which totally contains essential structured information about 150,000 papers on computer science.

4.1 Experimental Results

In Figure 2, we show OLIE's extraction results on instance accuracy and compared them with some typical algorithms (i.e. CRF and HMM). An obvious improvement is obtained due to two main reasons. First,

additional information such as vision and database is utilized to help the extraction. Second, the labeling process is based on elements instead of words.

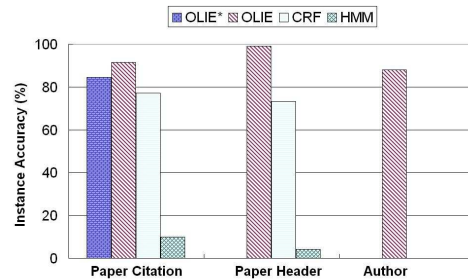


Figure 2. Instance accuracy by different algorithms

To test the effectiveness of using object elements instead of words, we discard database features during the extraction. The result is shown in Figure 2 corresponding to the OLIE* method. We can see that, though the result is not so satisfying as OLIE, an improvement is still obtained compared with CRF and HMM.

5 Conclusion

By leveraging the advantages of both Web IE and classic IE techniques, we propose an *Object-Level Information Extraction (OLIE)* approach by extending the *Conditional Random Fields (CRF)* algorithm with more Web-specific information such as vision features and database features. The novelty of this approach lies in that it utilizes as much available Web information as possible to assist the extraction process.

References

- [1] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.
- [2] B. Liu, R. Grossman, and Y. Zhai. Mining data records in web pages. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.
- [3] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: Bringing order to web objects. In *Proceedings of WWW Conference*, 2005.
- [4] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. 2d conditional random fields for web information extraction. In *Proceedings of ICML conference*, 2005.