

# Teaching Resources Information Extraction System Based on Web and XML

Wei CAO<sup>a</sup>, Zhiyi FANG<sup>b</sup>, Wei LI<sup>b</sup>, Peng XU<sup>b</sup>, Shuang CUI<sup>b</sup>

<sup>a</sup>College of Information Technology, Liaodong University, China

<sup>b</sup>School of Computer Science and Technology, Jilin University, China  
fangzy@jlu.edu.cn, musicalife@live.com

**Abstract:** This paper proposes a teaching resource information extraction system that based on Web pages. Using the semi-automatic method to extract the teaching resources data of designated websites, extraction results represent in XML format, and carry on the data with other systems or databases for data exchange.

**Keywords:** Teaching resources, Information extraction, XML

## Introduction

With the increase of websites, the teaching resources data present geometric progression to increase. At the same time, website teaching resources[1,2] exist with the form of Semi-structured data. These data are represented mostly through HTML[3] languages at present, and a prominent characteristic of HTML language is the structure is hidden, irregular or incomplete, which make WEB pages only suitable for people to read, but difficult to summarize analysis by the automatic machines. The emergence and development of Web information extraction[4] technology and XML[5] technology is making up for this deficiency. In this paper, the research for the technology of Web teaching resources information extraction is to build an extraction system of teaching resource information.

## 1. System Design

### 1.1 Establishment of Teaching Resources Semantic Model

We choose “limited XML” as the semantic model of the system, and take limited DTD express pattern. The node types in the “pattern tree” are atomic targets, set targets, meta-group targets and member targets. The establishment of teaching resources semantic model is according to the implicit semantic structure of the pages set, accessing subset from the description of metadata standard teaching resources elements, using limited XML to express the level of relations and establishing the appropriate pattern tree.

### 1.2 Extraction Rule and Extraction Knowledge

The extraction rule actually is a teaching resource attribute element object’s information extraction path. The extraction knowledge is corresponding with the teaching resources. It is

formed by the reasonable assembly of all the attribute element objects' extraction rule. Finally, extraction knowledge is used to carry on the extraction to the teaching resources. The work including: extraction rule production, extraction rule library and extraction knowledge library, teaching resources information extraction, extraction rule section access and optimization, and extraction rule assembly.

## 2. System evaluations

In the foundation of above principle, we have developed a teaching resources structure information extraction prototype system independently. This system divides into three constituents: extraction knowledge learning system, real-time information extraction system and user resources referral system. The prototype system uses development kit: Delphi7.0, database: SQL Server 2000 and the operating system is Windows2003.

We chose 3 websites page sets as the sample page sets to carry on the study, and to carry on the extraction experiment to the contained teaching resources. The result appraisal of the extraction test can see in Table 1.

Table 1. Effect of teaching resources extraction test

Name	Whether can extract	Study number of times	accuracy rate	recalling rate
K12 China secondary and elementary school education teaching web	Can	1	100%	99.1%
Secondary and elementary school education resources center	Can	1	99.6%	98.7%
Chinese teaching resources net	Can	1	100%	100%

From Table 1, we can see that our prototype system is quite ideal in actual extraction application effect. It can extract page's teaching resources basically. Extraction accuracy rate is between 99.6% to 100%, and recalling rate is between 98.7% to 100%. These websites' teaching resources are the entire cross pages. This also explained that our system to is also effective the cross page information extraction of teaching resources information method.

## 3. Conclusion

This paper proposes a teaching resource information extraction system that based on web pages. We use the semi-automatic method to extract the teaching resources data of designated websites. According to the model we construct a test system. The application effect has been quite ideal.

## References

- [1] D. Floresu, A. Levy, A.Mendelzon. (1998). Database Techniques for the Word-Wide Web:A Survey.ACM SIGMOD Record, 27(3).
- [2] P. Buneman. (1997). Semistructured Data. In Proceedings of the ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems. Tucson, Arizona, PP.117-121.
- [3] HTML 4.01 Specification. <http://www.w3.org/TR/htm14011>.
- [4] Gio Wiederhold. (1992). Mediators in the Architecture of Future Information Systems. IEEE Computer, 25(3): 38-49.
- [5] XML[EB/OL]. <http://www.w3c.org/XML>.