Study of Extraction for Web Pages Information Based on XML

Suming Li

School of National Education, Nanchang Institute of science & Technology, Nanchang, 330108, China.

niefengzju@163.com

Keywords: XML; web pages; information extraction; knowledge base.

Abstract. This paper proposes a web information platform based on XML. First, the information platform combines the advantages of existing different extraction technology, automatically extracts the key information in accordance with XML technology, next translates key information into structural and extensible XML documents, finally, concludes corresponding extraction rules by a group of similar pages, and then finishes the extraction for web pages information by these extraction rule.

1. Introduction

Web information extraction implements extraction for web information source. Its goal is to extract special factual information from semi-structural or non-structural information. For example, to extract details of terrorist incident the from news report, such as time, address, perpetrators, attack targets and used weapons, etc. [1, 2]. To extract new products information of from economic news, such as company name, product name, publishing time and production performance, etc. To extract the medical records from patients, such as symptoms, diagnostic records, test results and the prescriptions, etc.

2. The Study Status

The study of information extraction appeared in the early 20th century 90's. The study of foreign countries mainly included some aspects: Sergey Brin from Stanford proposes DIPRE algorithm, which can extract web data relationships [3]; N.Sundareson from the research Centre of IBM discussed double meaning problem of web document and proposed improved algorithms, in addition, dug abbreviation and full name of English word. There were some results of domestic research. Zhou from Fudan University in Shanghai studied mode extraction of semi-structural documents, and proposed incremental pattern mining algorithm. Zhang from Nanjing University constructed extraction of semi-structural data with OEM [4]. These researches of web extraction had further analyzed Internet data in accordance with the characteristic of semi-structure documents, and replaced the final result of information extraction with knowledge [5, 6].

3. The Study Contents

3.1 The General Description.

This paper focuses on studying how to extract data of interest to the user from semi-structural web pages, and attempts to propose a web information platform based on XML. The core of the platform is to generate extraction rules. Actually, the goal of the extraction rules is to position interesting information. To begin with, the samples web pages will be translated into XML document of well-structure. The area of interest to the user can be found from samples XML documents. Then finding out locating information of requirement extraction in the area, and implementing the inductive learning in accordance with of locating information of the different samples pages, so as to find out locating information of interesting information of these pages and construct extraction rules

based on XSLT document, finally, extracting the actual information by these extraction rules. The implementation process is shown in figure 1.

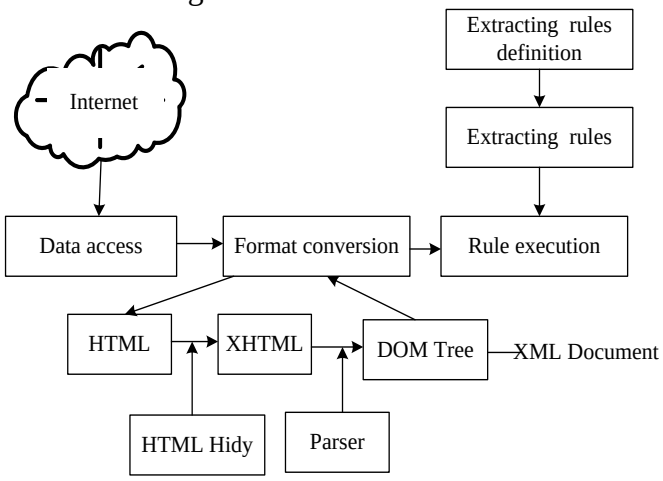


Figure.1 Extraction process

3.2 The Platform Goal.

The information platform combines the advantages of existing different extraction technology, automatically extracts the key information in accordance with XML technology, next translates key information into structural and extensible XML documents.

This paper wishes to conclude corresponding extraction rules by a group of similar pages, and then finishes the extraction for web pages information by these extraction rules.

3.3 The Basic Idea of Design.

Firstly, the platform acquires the sample web pages in accordance with user-specified URL, and translates the web pages into XHTML by HTML Tidy.

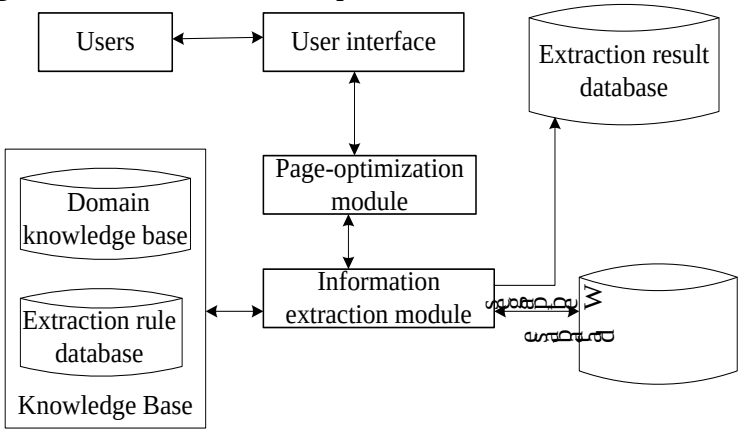
Secondly, XHTML document is parsed into DOM (Document Object Model) tree structure by XML Parser, and the DOM tree is referred to presentation of web pages within systems.

Finally, DOM tree structure is transformed into result XML document by XSLT.

3.4 The Overall Framework.

(1) Knowledge Base and Database.

The system base includes the knowledge base and database. The knowledge includes domain knowledge base and extraction rule database. The database includes extraction result database and web pages database. In practical operation, the construction of knowledge base and database from extraction system is complex, while this paper focus on information extraction, so will not delve into detail of the knowledge base and database. The platform overall framework is shown in figure 2.



Figrue.2 The platform overall framework

(2) Page-optimization Module.

The main function of page-optimization module mainly optimizes learning pages and extraction pages, so as to translate the unregulated or incomplete web page into well-formed XHTML document, and parse it into DOM structure tree.

(3) Information Extraction Module.

The information extraction module is the core of the platform. The information extraction is premised on acquiring the extraction rules, and the goal of any information extraction is to get reliable extraction rules, then extract information with extraction rules. Therefore, the operation process of information extraction is divided into two steps: first, implementing samples for study so as to acquire extraction rules, and then extracting information with extraction rules.

4. The Knowledge Base and Database of Platform

4.1 Constructing Domain Knowledge Base.

The function of domain knowledge base mainly includes the follow two factors:

(1) It must provide users with query navigation, so as to make operation more intelligent and more convenient for users. It works like this: some important webs URL are added into corresponding domain.

(2) It must provide rule management support in logic and methods. On way to store extraction rules in accordance with sub-domain.

The domain of this paper is professional website of publishing him same kinds of information, information of the domain knowledge base extraction includes the basic concept, attribute, entity and rule, etc. For example, for the kind of book information of publishing company, the domain knowledge base of it includes all kinds of the basic concept and attribute of books. This paper defines that every domain of domain knowledge base should form hierarchy with affiliations relationship, while the root is fictitious.

4.2 The Extraction Rules Base.

The platform adopts different rules for different domain and website, with the running of the system, many rules can be produced, which will be stored in the extraction rules base. When the system needs extract information, it first needs find whether there is a reusable rule from the rules base, if there is reusable rules, the corresponding rules will be extracted.

4.3 The Extraction Result Database and Web Page Database.

The final extraction result contains XML document of interest to the user, and these XML pages are stored into the extraction result database. The goal of Native XML database is to store XML document, which stores XML document with its format of XML document. Unlike the other database, the internal model is based on XML document format.

4.4 The Page-optimization Database.

(1) TIDY page document.

The main role of the TIDY page document is to fix web page and translated it into qualified XHTML document.

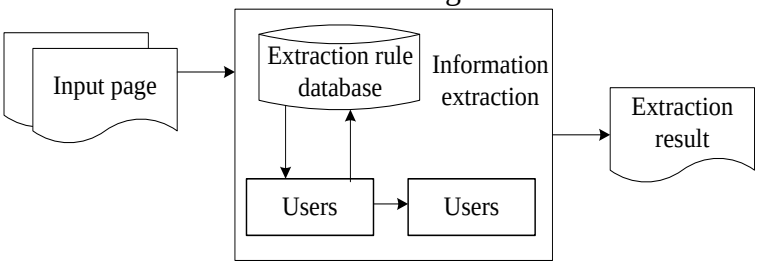
HTML Tidy is a powerful tool of open-source software, which can fix some common errors of HTML document and produce well-formed equivalent document. This paper integrates the class libraries of Tidy into the system. Web page is preprocessed by Tidy, and source HTML document translated into equivalent XHTML document.

(2) The Page Parser.

HTML DOM tree is described way of web page, which is established in accordance with HTML label of web page. It is tree structure of hierarchy, which every node is a single HTML element. Therefore, the path of hierarchy of DOM can be understood as “the coordinates” of extraction. The required information can be extracted with the acquirement of coordinates. In the process, XML document is loaded into memory and produces XML DOM tree, so as to produce rules based on DOM with extraction rule module.

4.5 The Page-optimization Database.

In the application of web information, using the wrapper to extract web information. In fact, the wrapper is software process, which extracts rules with defining information, and extracts information data from inputting web page. The extracting information will be translated into described information of special format, which will provide further study for the other information system. The work flowchart of information extraction is shown in figure 3.



Figrue.3. Work flowchart of information extraction

(1)The Basis for Rule Learning. The rule is also called as pattern in the different literature.

The core of Wrapper system is to extract rule. To construct accurate and robust extraction rule is a critical priority, also in pursuit of the goal for any extraction system.

This paper produce extraction rule by using contained structure feature, position feature, display feature, semantic feature and instance feature of HTML document. The steps of the rule learning are as follows: to begin with, determining sample page set; then implementing sample learning and producing the extraction rule.

(2) The Description of Information Extraction. When getting extraction rule XSLT document, constructing an information extraction Wrapper by implementing XLST.

5. Summary

This paper propose a method that it lay the favourable foundation for web information extraction, however, the scope of its applicability is limited. When encountering web page of complex structure and lack of semantics, the accurate ratio of extraction will be reduced. Therefore, the method of this paper needs to learn more about the extraction rule situablity and algotithm to slove the information complexity,in addition, needs to strengthen the authority and validity of information.

References

[1]. HU D D, MENG X F. An Automatic Extraction Method of Web Data Based on Tree Structure, Journal of Computer Research and Development, Vol. 9 (2010) No.20:77-78.

[2]. ZHANG S H, XU L H. Web Information Extraction Based on Sample Case, Journal of Henan University(Natural Science Edition), Vol. 11 (2012) No.22:17-19.

[3]. ZHOU J, ZHOU M. Automatic Extraction of Web Page Information Based on XML, Computer and Application, Vol. 11 (2010) No.11:21-23.

[4]. PENG H, CAI M L, CHEN J F. Visual Representation Model and Automatic Keywords Extraction Algorithm for Hub Web Pages, Journal of Computer Applications, Vol. 32(2012) No.8:2361-2364.

[5]. LIANG H, ZUO W L, REN F. Ontology Instance Based on Attributes Extracting for Deep Web, Journal of Chinese Computer Systems, Vol. 30(2009) No.5:8884-8887.

[6]. LIU W, YAN H L, XIAO J G. Solution for Automatic Web Review Extraction, Journal of Software, Vol. 21(2010) No. 12:3220-3236.