Research on the Application of Web Information Extraction Based On Semi Structured XML

Guo-Jun YANG
Neijiang Normal University
404510815@qq.com

Keywords: Semi Structured, Information Extraction, Extraction Rule, Framework Design.

Abstract. This article in view of the present a lot of Web data information, Starting from the research background of information extraction, Detailed analysis of various forms of Web information, Presented a rule based on information extraction, Basic objectives, design ideas, general framework and main contents.

Research Background

With the rapid develop of Internet, The Web has evolved into a huge, distribution and sharing of information resources, But most of the Web data appear in the form of HTML, Lack of description of the data itself, Do not contain clear semantic information, Mode is not clear, The structure is also not good. This makes the application cannot parse directly and use the huge amounts of information on the Web, In order to enhance the usability of Web data, the Web information extraction technology, which if through packing existing Web information sources, information on a Web page in a more structured way of extracted, using the data in the Web offers the potential for the application.

Research Status

The study of information originated from the beginning of 1990s, Study abroad in the following aspects: Sergey Brin at Stanford University put forward that DIPRE algorithm can be found on Web document data relations; IBM research center, N. Sundaresan with others people have discussed and proposed the improving algorithm on the problem of double meaning in Web documents, and the English words abbreviations and full name of the Web were excavated, The researches in China: Fudan University, Zhou Aoying and others have studied the semi structured document model extraction, who also proposed the incremental pattern mining algorithm; Zhang Fuyan of Nanjing University used OEM model to construct the semi structured data extractor. These Web extraction studies used the characteristics of semi structured documents for the characteristics of the data on the Internet to search and analysis deeply, with knowledge instead of information as the final results of information access.

What is Web Information Extractions

Web Information Extraction(Web Information Extraction, referred to as WebIE is using the Web as the information source of a class of information to be extracted. who's main purpose is that extracts specific factual information from semi structure or non structure information. For
example, from the news reports extract details of terrorist incidents: time, position, crimes, victims, targets, the use of the weapons etc; Economic news from the company to release the circumstance of new products: company name, product name, release time, product performance etc; from the patient's medical record extract symptoms, diagnostic records, test results, prescriptions and so on, or direct extraction of information in a sentence or a word, etc.

Research Contents

General Description

This article focuses on how to extract the data which users are interested in from the semi-structured Web page, and tries to put forward a Web information extraction platform based on XML. The core of the work is to generate extraction rules. Here, the extraction rule is actually the location information of the information point of interest. First need to sample web pages into well structured XML documents; By researching the area of the user's interest from the sample XML document; And granularity find the location information of the specific information to be extracted in this area; Then, summarized the location information of different sample page of learning, learning the interested information point location information of the page, and constructing the rules is extracted by XSLT document representation. Finally, the extraction rules are applied to extract the actual information. Information extraction flow as to

![Information Extraction Flow Chart](image)

**Figure 1. Information Extraction Flow Chart.**

Platform Target

The purpose of this platform of information extraction is combined with the advantages of existing different extraction technology, based on the XML technology, automatically extract key information from the Web page, and be structured, strong extensibility of XML documents. General structure diagram as to Figure 2.

In this paper, we hope that through a similar set of pages, can be summed up the corresponding extraction rules, and then use the extraction rules to extract page information.
The Basic Idea of Design

This system according to the user specified URL to obtain the sample page data and this web page uses HTML Tidy to convert to XHTML.

Then use the XML Parser to parse the XHTML document into DOM tree structure, so that the DOM tree will become a Web pages' form of expression in the system.

Finally, on the basis of obtaining the DOM tree, use XSLT to transform the DOM tree structure to the result of XML documents.

Figure 2. General Structure Diagram.

General Framework of Platform

Knowledge Base and Database. The library in the system includes Knowledge Base and database, Knowledge Base includes domain Knowledge Base and extracting rules library. The database includes the result and the Web page database.

In practice, to extract the construction of Knowledge Base and database in the system is more complex, and in the relevant literature [48], this problem has been discussed in this paper. The focus of this paper is the research of information extraction, so the part about the Knowledge Base and database do not do in-depth exposition. General structure as to figure 2.

Page Optimization Module. The main needle are to be studied pages and to be extracted pages to be optimized, so that the structure of incomplete of irregular Web pages into a good structure of XHTML document, and parse into a DOM tree structure.

Information Extraction Module. Information extraction is the core of this paper, information extraction to obtain extracted rules as a precondition for any information extraction research are committed to obtain extraction rules with robust and reliable, and then use the extraction rules for information extraction. Therefore, this part is divided into two steps: first, a
sample study carried out in order to obtain extraction rules; then use the rules for information extraction.

**Knowledge Base and Database in Platform**

**Construction Domain Knowledge Base.** The function of domain Knowledge Base mainly includes the following points:

1) To provide users with query navigation services, the users cannot go so far helpless when they begin use it. The method is to add some of the important sites URL to the under of corresponding area.

2) To provide logical and method support for rule management, the method is to store the extraction rules according to the sub classification.

The field of this article refers to the professional web site of the same kind of information, Domain knowledge base is to extract the information field contains the basic concepts, attributes, entities, rule and so on. Such as press publishing web site is a kind of information about the books, it's domain knowledge base should contain all kinds of basic concepts and attributes, such as the requirements of the book. In this paper, the domain knowledge base is formed by a hierarchy tree, and the root is virtual who also can be referred to as the “root”.

**Extract Rule Base.** Extraction rules repository storage is already learning to extract rules, extracting rules is to extract knowledge identification mode. Different areas and sites are used ion different rules, as the extraction system running, many rules will be produced, the system naturally requires a library to store these rules. When the system needs to carry on the information extraction, first of all to the rules in the library to find whether there is can be repeated use of the rules, if you can directly extract the corresponding rules from the repository, and you do not have to regenerate new rules for similar websites or web pages.

**Extraction Result Database and Web Page Database.** In this article, the results are extracted from the XML document containing the information point of interest of the user, the extracted results database is stored in these XML pages. Native XML DataBase, also known as the origin of the XML database which is specially designed for storing XML documents in the database, it is in form of XML document itself to store XML documents, and other databases of different is its internal model which is based on the XML document format.

Deposit page database from a Web page to retrieve the source document. This part of the document has been processed by system, including cleaned and repaired and the marked page file.

**Page Optimization Module**

**Tidy Page Document.** This part of the need is going to repair the Web page to be converted into a standard XHTML document, known as the TIDY in this paper.

HTML Tidy is a powerful tool for open source code, which can be used to correct the common errors in HTML documents and generate a good equivalent document format. This article uses the Tidy library, which is integrated into the system. Web page will be pretreated by Tidy, make the source HTML document into equivalent XHTML document. XHTML format conversion flow as to Figure 3.
Page Parser. The HTML DOM tree is a description of the HTML page, which is based on the meaning of the Web page HTML tag and set up a hierarchy tree structure, whose each node is a single HTML element. Therefore, in this paper, the path of the DOM hierarchy is understood as the "coordinate, and the information obtained through the acquisition and understanding of the coordinates obtained. In this process, the XML file is loaded into the memory to generate XML DOM tree, which can be used to extract the rule learning module to generate the rules based DOM. Page tree structure diagram as to Figure 4.

Information Extraction Module

In the application of Web information, we use a wrapper for information extraction. Wrapper is a software process, applying the information extraction rules which have been defined, will show in the input information in the Web page data extract, transform into information
described in a specific format, available to other information systems to do further research and information extraction in the working process below. Flow chart of information extraction as to Figure 5.

**Basis of rule learning.** In the application of Web information, we use a wrapper for information extraction. Wrapper is a software process, applying the information extraction rules which have been defined, will show in the input information in the Web page data extract, transform into information described in a specific format, available to other information systems to do further research and information extraction in the working process below. Flow chart of information extraction as to Figure 5.

Rule, which is also called the pattern in the literature. The core of the Wrapper is to extract rules, construction accurate robust extraction rule is a priority, which is the goal of any extraction system.

This article mainly uses the structural features, location features, display characteristics, semantic features and reference form extraction rules of HTML.

![Figure 5. Flow Chart of Information Extraction.](image)

**Rule Learning Step.** Determine the sample page set Sample learning to generate extraction rules

**Description of the Process of Information Extraction.** After obtained the extraction rule XSLT documents, to construct a wrapper for information extraction only needs to perform the XSLT. Extraction rule chart as to Figure 6.

**Expectation**

In this paper, the applicable scope of the proposed method is still limited. When the page structure is relatively complex and lack of the semantic meaning, the accuracy of the extraction will be reduced. Therefore, the next goal is to strengthen the adaptability of learning extraction rules.

Source of authority and effectiveness affect the accuracy of information extraction, and the collection of the target document is also a problem to be solved. This article also requires an algorithm to determine the authority of Web pages, which is another aspect of the need for further research.
In this paper, the applicable scope of the proposed method is still limited. When the page structure is relatively complex and lacks semantic meaning, the accuracy of the extraction will be reduced. Therefore, the next goal is to strengthen the adaptability of learning extraction rules.

Source of authority and effectiveness affect the accuracy of information extraction, and the collection of the target document is also a problem to be solved. This article also requires an algorithm to determine the authority of Web pages, which is another aspect of the need for further research.

References

