Analysis of the HTML to XML Conversion Method

LI Busheng1, a, HU Jingfang1, b
1School of Information Engineering, Jingdezhen Ceramic Institute, Jiangxi, P.R. China
aemail: abulbs@163.com, bemail:jdzhjf@163.com

Abstract

In this paper, the features of HTML and XML were compared, and expounds the necessity of transition from HTML to XML, and finally introduces document conversion, XHTML conversion and intelligent instead of three kinds of HTML to XML conversion method.

Keyword: HTML； XML； conversion method； XHTML

Characteristics of HTML

HTML (Hypertext Markup Language) is a universal language for creating Webpage and to release information, it is YISHION text stored in the form of organization, to label definition document. Provided the cross plat form file sharing. In the HTML document, can be embedded in other objects, such as electronic forms, video, audio and various applications and so on, through the uniform resource locator can realize the hypertext links between Web nodes. HTML has the following features, format and grammar is relatively simple, easy to learn, will the data with some control markers, even if no programming experience can easily use HTML to design Webpage; and all of the control tag HTML is fixed [1], the number is limited, provides functions and related properties the setting is fixed, easy to remember; rules more flexible, such as the control of English marker size marker to write in no difference. In addition, the control flag must have the end tag corresponding to no strict requirements. Need the simplicity of HTML is more suitable for low cost information publishing; HTML as Web common information description method of strong universality, can realize different platform document sharing; to create more flexible, HTML document is a plain text file, can use a variety of editing tools for creating. The major disadvantage of HTML is:

(1) Performance is too simple.

(2) Link easily broken, chain destination address changes, chain source cannot automatically correct.

(3) The flowers during the retrieval time are longer, the retrieved content targeted poor, and many returned results.

(4) Poor scalability, HTML tag set is fixed, not to allow users to define their

own identity.

(5) The lack of semantics, HTML is a marker, it will not reveal the nature of the information content, and the computer cannot know the exact meaning of each section of text.

HTML and XML are used for the network information organization and communication, are in the form of texts written in storage, and are structured information based on international standards [2]. But HTML only shows data seems to be what kind of, and the XML is that data is what mean. Using XML can create their own tags, these tags can more accurately describe the user what they want, but HTML can not be used to define a new application, and this is the biggest difference between XML and HTML. XML has overcome some limitations of the HTML, has the broad application prospect.

The necessity of transformation

Most of the current Webpage still is mainly composed of HTML. HTML of the inherent shortcomings of the original network information organization mode can't meet the requirements of the development of the new. Because the relevant development of the XML technology continues to mature, more and more websites gradually in XML design. In this process, not only will the new content is stored in XML format and transmission, but also consider compatible with the original data, so it is necessary to convert existing HTML Webpage into XML data form more flexible processing and application. At the same time, to the previous accumulated HTML documents continue to play a role in the new environment, the XML conversion is undoubtedly a solution. HTML to XML conversion is helpful for Webpage information integration, extraction, retrieval, filtering or mining analysis.

(1) Facilitate the information integration of heterogeneous system under Network Environment

In the network environment, because of the existence of the system platform and database operation of heterogeneous, leading to the exchange of information and work with difficulty. One of the data exchange between heterogeneous systems is to adopt a unified information exchange format. XML because of its custom and scalability advantages [3], which is convenient for expressing various types of data between heterogeneous databases, and can be used as middleware, unify the data interface, convenient for the exchange of information between different databases and colleagues. XML can be used to construct the data layer integration, will transform the source data into the data integration, simplification of integration system query translation, query mechanisms provide unified multi data source for the user, in a unified way using a variety of data from different sources of data, different shielding each data source in the structure, running environment on the.

(2) The organization and management for network information

With the rapid increase of content and application of Web information, to realize the automation of network information management is very necessary. Through the HTML interactive operation is often due to the service provider to change the layout, add new content or change the URL, the browser error, and change the need to spend a lot of time. XML can provide a comprehensive service through the browser defined by XML for data description and interface format, exchange and catalog information and updated automatically, automatic classification processing to achieve the directory. Because of the information resources of the growth rate, HTML to describe unstructured data on the lack of adequate capacity, and any through artificial network information resources organization practices are not feasible, metadata is generated for the contradiction between the efficiency to solve the problem of enhancing network information resources organization and arrangement of the order, but it is difficult for all the metadata of network information resources the objective description, and the use of XML tag data with semantic, can increase their exchange in different network systems flexibility and understanding degree greatly, thus to solve the problems of organization and management of network information.

(3) For the convenience of network information retrieval and filtering

Because the XML marking a clear expression of its meaning, the search engine will be able to accurately locate specific information according to the relationship between keywords and content, which according to the keywords provided by the user, clearly know the semantic user expression and returns the correct results. Semantic structured XML can be used as the standard for the exchange of structured data, will improve the retrieval results. HTML cannot know one article titles, authors, abstract, conclusion, and XML can provide information about these structures. HTML to complete a query process is usually more complex, but with XML it can automatically complete a series of query process. XML makes the agent becomes more personalized information retrieval. The use of XML can also be found and positioning of the network bad information, so as to take effective measures to filter and purify the network environment.

(4) To facilitate the implementation of network data mining and knowledge discovery

Extracting information from HTML documents is often difficult, but XML has the semantics, facilitate information extraction and analysis. XML enables data from different sources easily together; to retrieve multiple incompatible databases as possible, thereby mining brings a new way to solve the network data. Scalability and flexibility of XML allows description of different kinds of data in the application software, which can collect the data records". At the same time, because the XML data is self describing, the label has the semantics; data does not need to have internal description can be exchange and processing. Therefore, XML can be regarded as a kind of semi structured data model, can be easily described the document and relation database in XML attribute correspondence, information query and extraction.

The main method of conversion

1 HTML documents directly to XML conversion

XML is a relatively strict markup language, it requires a document all statements in full compliance with the XML specification, as long as the XML document has a little error, the XML application can not correctly handle. Therefore, to convert HTML to XML, the first first is added to the HTML XML statement, tell the client processing program, this is a XML document and related resources in what place; second is to check the HTML document has no grammatical errors, if any, should be carried out strictly according to the design specification of HTML correction; again is to define entities, unlike the HTML, all the entities in XML must first define after use, such as some graphics file HTML document used to that entity, in XML must first be defined to use. In the HTML document to XML document automatic conversion process, requirements well formed HTML documents, i.e. end-to-end mark all elements must be paired, a nested hierarchy of all elements must be properly, all attribute values in double quotes ("symbol")form, all from the elements to </> end. The conversion process is mainly to solve the problem of HTML document and its mode of information sets to be expressed, can be automatically selected through the program, and the use of manual for further processing of HTML text, to model the omission of information to be added; thus forming a complete data model. According to the extracted pattern, determine the semantic relation attribute name and object the object between the specified markups to be converted; the information in the HTML document, the corresponding relations and clear documentation tags these are converted to HTML documents and XML; according to these relations, scanning HTML document and output the corresponding XML results.

2 Using XHTML conversions

To migrate HTML to XML, this can be retrofitted to HTML document using XHTML. XHTML combines the advantages of HTML and XML, since it is closer to HTML, so you can easily use HTML modified and simplified, the formation of a new XHTML document, implement the HTML to XML transition. In the conversion process should pay attention to several issues: first, the size of XHTML to write marker sensitive, therefore, in the HTML to enhance the readability of the document some skills in the XHTML will not apply, as defined in the HTML element attributes use uppercase characters, while the specific numerical can use lowercase, readable will some strong, but is defined in the XHTML element attributes must use lowercase. Secondly, XHTML strict elements must be to mark start and end markers. In HTML is not strictly in accordance with the use of at the beginning of each section and in the end the use of , but in XHTML of all non empty elements are required to be closed. Again, all the XHTML attributes are required to use quotation marks to indicate, for example HTML <table border = 2> statements need to be rewritten as <table border = "2" > in XHTML. In addition, like the <head> and <body> elements in

HTML can be omitted, and in XHTML is required.

3 The use of intelligent agents for conversion

Intelligent agent is usually refers to simulate human behavior, according to the perception of the environment of independent operation and to provide the corresponding service program. In the Internet category can be defined as the proxy user or other program under the environment of network, complete the relevant operation software entities in an active way [4]. Intelligent agents can realize the purpose of related plans and create, timing and interactive execution, respond to changes in the network environment. Intelligent agent has the characteristics of autonomy, agent, intelligent, interactive, mobility and adaptability, can automatically acquire, analyze and process the data in the user does not need to intervene in cases; these technical characteristics has been fully applied in the network information organization management. Intelligent agent is able to identify web features, can automatically search the relevant content, and then converted into the corresponding XML format. Through the retrieval intelligent agent find out about RDF in the news information resources on the Internet, including metadata, path, the article identification and structural rules, then according to the structure of the article with the XML tag said the title of the article, author, date, location, location and other elements, the non structure HTML is automatically converted to include many structured XML the computer to identify the information processing, well on this basis can content retrieval by XML label.

Conclusions

HTML and XML are used for the network information organization and communication [5], are in the form of texts written in storage, and are structured information based on international standards. But HTML only shows data seems to be what kind of, and the XML is that data is what mean. Using XML can create their own tags, these tags can more accurately describe the user what they want, but HTML can not be used to define a new application, and this is the biggest difference between XML and HTML. XML has overcome some limitations of the HTML, has the broad application prospect.

Reference

[1] (U) YI Wien. Wang Chunnan, Liu Yongjin. XML advanced programming. Tsinghua University press,.2009.2

[2] Chu Jiu Liang ed., Web front-end development technologies: HTML\CSS\JavaScript, Tsinghua University press, 2013

[3] As Ling, Li Shengtao, Cheng Xueqi. XML database information exchange mechanism based on.2003. Computer engineering and Applications

[4] Tan Hanhua.XML language and built on the base of the XML three layer

C/S model application. Coastal enterprises and technology.2008.5
[5] Luo Yan. Research on technology of heterogeneous exchange based on XML: (Master thesis). Nanchang: Nanchang University, 2011