Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 3170 – 3174

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Research and Design on E-government Information Retrieval Model

Xiaoxing Liu[a*], Changxia Hu[a]

[a] *School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China*

Abstract

At present, most government websites have not given full play to the internet's characteristic such as share, interaction. The public satisfaction is poor. Therefore, this article puts forward an e-government information retrieval model based on the metadata. Firstly it builds an e-government information resource database, and all kinds of different types of resources are uniformly stored in the resource database; Secondly, the resources are described by metadata, and e-government information resources metadata scheme is designed; Thirdly, the metadata extraction algorithm is designed to extract the metadata elements of e-government information resources; Finally, e-government information resources storage and retrieval schemes are designed. The experiment result shows that the information retrieval method based on metadata can improve the e-government information retrieval efficiency and precision, and realizes the share of e-government resources. Thus it is a practical approach for the task of e-government information retrieval.

Keywords: E-government; Metadata; Chinese lexical analysis; Information retrieval model; Dublin Core

1. Introduction

E-government is a compact, efficient government operation mode which uses the modern information technology in its management and service functions. It realizes the restructuring optimization of the organization structure and the working process, and surpasses the restriction time, space and department

[*] * Corresponding author.
E-mail address: liuxx@stdu.edu.cn

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.460

division [1]. After the phase of office automation, three gold project and government online project, e-government has entered the system integration and information resource integration phase.

The government information resources are those all produced in government interior or although produced outside the government but affect every government business activities. The government is the largest social information resource owner, producer, user and transporter, and plays a leading role to the social information resource development and utilization. Its service objects include government, enterprise and public.

The government information resources have the following characteristics: amount of information is large, wide range and the higher requirements to information accuracy and security. By the policy effect, the information store space is widely distributed, relatively independent in scattered government departments. The storage mechanism differs in thousands ways, and there are a lot of overlap. Thus information islands are formed, and resources cannot be fully shared.

To solve these problems above, the paper puts forward an e-government information retrieval model based on the metadata. Firstly it builds an e-government information resource database, and all kinds of different types of resources are uniformly stored in the resource database; Secondly, the resources are described by metadata, and e-government information resources metadata scheme is designed; Thirdly, the metadata extraction algorithm is designed to extract the metadata elements of e-government information resources; Finally, e-government information resources storage and retrieval schemes are designed. The design process will be expatiated in the next sections.

2. Metadata model design

2.1. Introduction to metadata

Metadata is the data on the data, this term refers to any data used to help identify, locate and describe electronic network resources. Metadata describes the attributes, types and relationship of resources [2]. It changes the information on the web from unreliable to reliable, from not modified by the semantic meaning to modified, and provides a premise and possibility for the full development and utilization of web resources.

In this paper, the object described by metadata is e-government information resource. The e-government information resource metadata is designed to promote e-government resources' classification, storage and retrieval. Its use can be summarized as follows:

(1) Summarize the meaning of the data.

(2) Allow users to manage, find, access to data.

(3) Help users understand the data, and determine whether the data meet their needs.

(4) Provide a consistent description of the data, and promote the sharing and exchange of data.

(5) Create other applications based on the metadata.

The metadata provides an effective way to find, manage and share data. At the same time, it has also brought another problem. The resources cannot be shared effectively with the non-uniform approach of metadata use, which causes the iterative development of resources. Therefore we need to define a common e-government resources metadata model, so that it can be widely adopted. On this basis, efficient finding, location, evaluation, acquisition and management can be achieved. The above-mentioned is the purpose of developing e-government resources metadata standard.

2.2. Metadata model designed by the system

Through the study of present e-government information metadata standard, twelve metadata elements are designed in the system, including title, creator, subject, identifier, date, publisher, language, assessment, alternative title, institution and contact. Each element is defined in great detailed, and the mark rule and use of each element are designed.

In practice, in order to simplify the design difficulty, some of elements can be chosen to be "the core data element", the other are "the optional data element". In this way, the whole model can be divided into two parts: core collection and optional element collection, among which title, creator, subject, identifier and date are core elements. And the rest are optional. With the five compulsive elements, a resource entity can be located precisely. When the user makes query request, he can find the information he needed quickly and precisely in the metadata database.

3. E-government information retrieval model based on metadata

E-government information retrieval model based on metadata is essentially a kind of metadata search engine. It retrieves metadata in a specific information resource library, and it includes web page collection system, metadata extraction system and information maintenance system [3, 4].

Web page collection system is responsible for the collection, storage of the information resources about particular subjects.

Metadata extraction system extracts the core metadata from the web page collected by web page collection system according to the metadata standard.

Information maintenance system is responsible for the release, query and maintenance of metadata resource library.

This paper designs a metadata information retrieval model in e-government system, as shown in Fig 1.

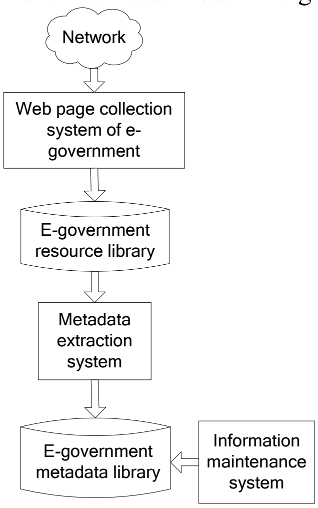
Network
Web page collection system of e-government
E-government resource library
Metadata extraction system
E-government metadata library
Information maintenance system

Fig. 1. The e-government information retrieval model based on metadata

First, through the web collection system of e-government we collect web resources related to the e-government subject relying on the subject library, and store the contents searched to the e-government resource library. Secondly, we extract metadata elements from collected web pages. The metadata information is stored in e-government resources library. When the user sends a query request to the system, the system will give the query request expanded to metadata database, and returns the query results to the user. Metadata extraction is the key in the whole model, which determines the speed and accuracy of information retrieval.

Metadata extraction includes the following four steps.

3.1. Web information collection

In order to collect some specific web information, a web spider program is designed to capture web pages. It looks for web pages through its link addresses, from a certain page (usually the first page)of the web site, reads the content of the web page, finds other link addresses in the web page, and then looks for the next page, and repeats the same cycle until capture all the pages of the web site. The Web Spider extracts the information like web title, URL, date of creation in the process of searching.

3.2. Web text information extraction

Metadata information search engine creates a web page index, and its object is text files. To the web spider, the captured web pages includes all kinds of files, such as html, photo, doc, PDF, multi-media, dynamic web page and so on. After these files are captured, all the text information will be extracted.

The system adopts regular expressions to remove the html labels, and extracts the text information in the web page. A regular expression is a formula that matches a kind of character string in a certain mode. The process of text information extraction is as follows:

(1) Obtain the source code of the web page (innerHTML attribute);

(2) Remove the specified html labels through regular expressions as follows,

```
o.innerHTML=o.innerHTML.replace(/(<\/?(?!br|p|img)[^>\/]*)\/?>/gi,'').
```

3.3. Web page subject extraction

Web page subject extraction includes the following steps:

(1) Make Chinese lexical analysis of the extracted text information, and label the part of speech.

Chinese lexical analysis is the basis of Chinese text information processing, directly related to the precision of the extracted subjects. Here we use the open source code of Chinese lexical analysis system of Chinese Academy of Sciences [5]. Its basic idea is to adopt the HMM model to make a Chinese word segmentation chart. In the stage of coarse segmentation, first get N results of the greatest probability, and then identify the words that are not logged in with the method of role labeling, and calculate the probability, add these words to the word segmentation chart, thereafter treat them as common words, and finally make dynamic planning to choose N results of the greatest probability. In practice, use java native interface to achieve the call of java in ICTCLAS system.

(2) Remove the part of speech of the stoplist and non-keywords.

Stoplist refers to the words that often appear in the text but without much actual meaning, mainly including empty words, adverbs, words of mood, like "also", "but", etc. The system uses the stoplist table of information retrieval laboratory of harbin institute of technology. Remove the stoplist from the extracted html text, at the same time, remove numerals, words of measure, adverbs, prepositions, conjunctions, auxiliary words, exclamations, adjectives from the text, and only keep the possible subjects.

(3) Calculate the word frequency, extract subjects.

In order to extract subjects from the text, first calculate how many times each word appears. By experience, the more times the words appear (not including part of speech of non-subjects), the more they can represent the main topic of the article. Thus according to the result of the word frequency, find N words of greatest frequency, and they are the N subjects of the article.

3.4. Metadata storage and retrieval

The system adopts XML as the e-government information describing language, which is widely used in the applications that follow Dublin Core. Through XML binding of e-government information metadata, the nesting relationship of the data elements in the intrinsic model is transformed into the nesting relationship of XML elements. This binding mode is clear and explicit. Through the increasingly universal sign language, achieve the mutual operation between metadata examples and resource library to the greatest extent. The storage of XML binding metadata can adopt the conventional relational database, and can also use recently developed XML database and updated RDF storage technology. Considering that XML database supports the system model better, adopt Tamino database to store and manage XML files, and the server is based on the platform. With JSP, Servlet, JavaBean and EJB technology, develop an actual application system through the programming interface provided by Tamino database.

4. Experiment and result analysis

The information retrieval model based on metadata designed in the paper mainly used in e-government information resource search. First through the web spider program to collect related web pages, and extract the core metadata in the web pages in accordance with the metadata model designed in the paper. Then establish metadata information database and provide the service of query in Tamino database.

The experiment firstly collects 800 web pages on e-government through spider program, and then runs the automatic metadata extraction program. The result is shown in Table 1.

Table 1. Core metadata extraction result

Metadata name	Extraction number	Right number	Accuracy Rate
Title	800	702	88%
Creator	800	603	75%
Subject	800	723	90%
Identifier	800	785	98%
Date	800	774	97%

As can be seen from the experimental result, information search accuracy rate of metadata-based search engine depends largely on the accuracy of metadata extraction. The metadata identifier and date are extracted by network spider program, accuracy reaches 95% above; The title metadata is extracted by <title> marker, however, some web pages do not fill correct title in <title> marker, 88% is correct; The subject metadata extraction depends on the accuracy of the above-mentioned segmentation system, 90% is right; The creator metadata has a lower accuracy rate of extraction, 75% is correct. The reason is that the extraction algorithm of author metadata in the system can only extract the fairly standard web pages, such as the format for the ‘author Tom’, etc. Compared with the current mainstream search engine, recall rate and precision of the system have been greatly improved.

References

[1] Sang M Lee, Xin Tan, Silvana Trimi. Current practices of leading E-government countries. Communications of the ACM, 2005, 48(10), pp. 99-104.

[2] Wu Jianzhong. DC metadata. Shanghai: Shanghai Science and Technology Literature Publishing House, 2000, vol. 10, pp. 50-51.

[3] Wang Fang, Word Segmentation Method Research Based on the Parse of XMI Information Retrieval. Journal of Information, 2008, vol. 1, pp. 121-123.

[4] Kobayashi M, Takeda K, Information Retrieval on the Web. ACM Computing Surveys, 2000, 32(2), pp. 144-173.

[5] Liu Qun, Zhang HuaPing, Yu HongKui. Chinese Lexical Analysis Using Cascaded Hidden Markov Model. Journal of computer research and development, 2004, 41(8), pp. 1421-1423.