Design and Development of the automated information gathering and issue system based on the small and medium-sized website

Zhao Ming[1], Zhang XiaoShuan[1], Ma Yong-ling[1], Zhu Zi-li[2], Zhao Xiao-jun[2]
[1](College of Information and Electronically Engineering, China Agricultural University, 100083, Beijing, China)
[2](WebCate Technologies Ins., Beijing 100089)

Abstract: - Along with the gradually applied and the popularization of the agriculture information, most of the counties and the cities have established small and medium-sized website faced to the agricultural and the country service in china, but quite a lot of these websites have the problems of few service contents, renewing slowly, lacking the means of gaining real-time information on-line automatically and so on. This article discussed the realization of an information automatic gathering and issue system which suits to the small and medium-sized website. This system could automatic gather and classify the website information, and carry on the essential support to the information searching and the browsing.
Key-Words: - Website; xml; information gathering; issuance system

1 Introduction

Along with the gradually applied and the popularization of the agricultural information, most of the counties and the cities have established small and medium-sized website faced to the agriculture and the country service in china, but the level of these websites are quite low mostly, which mainly problems in several aspects:

(1) The contents are not only few but also renewed slowly in Website.

(2) Although has the useful website information sources, it lacks the effective means and software to gain them.

(3) It can not manage the on-line information gained from the website effectively, which including the classification, storage and searches of the information.

Through the analysis, the article considered there were mainly two aspects which had brought these questions:

(1) Most of the companies in the country are shortage of the well-trained person on computer network. therefore, there are many problems on technical maintenance for the website after they are set up.

(2) If the website wants to be renew and provide the most recent and entire information, it has to face the enormous information collection and maintenance work, so the renewal and maintenance of the website content have already substitutes for the technical maintenance of the website and become the biggest bottleneck question in the website construction and the developing process [1-2].

In view of the existence problems on website above, we have developed an information automatic gathering and issuing system---WebCPS (Web Content Parse System). This set of system is mainly used to automatic collect the information, maintain the website resources, renew the website content and issue the information in time and comprehensively.

The characteristic of this gathering system is that it can track and gather the important websites about the agricultural in and out home real-time, and manage the information intelligently, which can provide the information for the decision-maker and the producer.

Using this system, we have established a agriculture information website for Quzhou of Hebei province in china, which not only provides the science and technology, the education, the market and the policy information for the decision-maker, the agricultural technique promoted personnel and the farmers, but also constructs a platform where people can exchange information on-line.

2 The design of the information automatic gathering and issuance system

2.1 The design principle of the system

The traditional gathering information technology mainly uses the model of "network spider + full text search"[3-5], and there are obvious drawbacks in this kind of model:

(1) It cannot gather the websites which adopt the technical that could display complex webpage (this kind of site takes up approximately all the Websites).

(2)It wastes network bandwidth and the storage space of the user

(3)It cannot realize the complex processing complete automatically to the webpage data

(4)It will lose the original logical relations after the data gathered

In view of the "network spider + full text research" model, we have proposed "the template defined and directional gathering model". This model realized the directional gathering to the pages on the base of the template defined, and its design idea is "the entire digitization" to the goal webpage by the XML data format firstly, and then separate the goal elements by the user's demand definition in the template. This is from the primitive webpage to the final information gathering.

The localization technology principle of webpage element digitization is the fully use of the structure characteristic of the webpage information in brief. When locates and extracts the data by the coordinate's way, there are big difference between webpage data and the common text data.

Firstly, the webpage data is the structured or semi-structured data, the information content is divided by each kind of HTML mark into different parts, and thus the data in the webpage will carry the accurate positional information itself.

Secondly, most of the webpage data is the HTML mark which has nothing to do with the content information, these marks will be disposed as the redundant data if use the traditional character matching filter model. The quantity of this kind of redundant data is extremely huge, and always occupies 80-90% of the entire webpage.

Thirdly, the webpage data is one kind of dynamic change data form (the webpage structure changes frequently), and the character processing of the traditional model cannot satisfy the request to the processing of the dynamic change data.

2.2 Gathering information by the structure

In traditional significance, the information gathering actually is the webpage downloading, and regards the HTML source document (or the HTML source document after filtering) as the text to carry on the full text index, and carries on the off-line browsing through the result of the full text research. This procedure can satisfy the user's keywords researching request, so the search engines are realized generally by this procedure [6-7].

But in the information gathering fields, the gathering information result is not only for searching, but also must be manage. Such as the information edition, the information statistics, and the storage information into the database. The" search engine" model above mention is unable to satisfy the information processing request well, because the plain text processing way of the information means it only can carry on each kind of text operation to the information. But the information after the WebCPS system processing, is one kind of structure information that each piece of information contains many fields and can distinguish the data type of the field is the text type, the numeric type or the date type.

2.3 The new characteristic brought by the structurize of the gathering result

The structure information that through WebCPS obtains, has provided the powerful support for the following processing of the information, but not only display the primitive webpage content and search by the keywords. Here lists the information processing method which the structure information can support:

(1) Carries on the index according to the information field, but not only can carry on the index to the entire information text; Its result is that the inquiry also can according to the information field;

(2) Carries on the rank, the subsection and the range inquiry operation to the information fields of the time and numeric type. For example it can carry on the price relations operation to the commodity information which is gathered;

(3) Carries on each kind of classification according to the information field to the information, for example classifies according to the author, or the origin of information;

(4) After the information is transformed, it can be input into the database as other information systems.

In the gathering process, the information which already gathers may affect the following gathering process; this is the important characteristic of the real-time dynamic alternate WebCPS system. This kind of real-time dynamic alternation is the difference between WebCPS and the traditional gathering system, a complete information gathering flow of WebCPS contains six steps to realize, as the figure 1 shows:

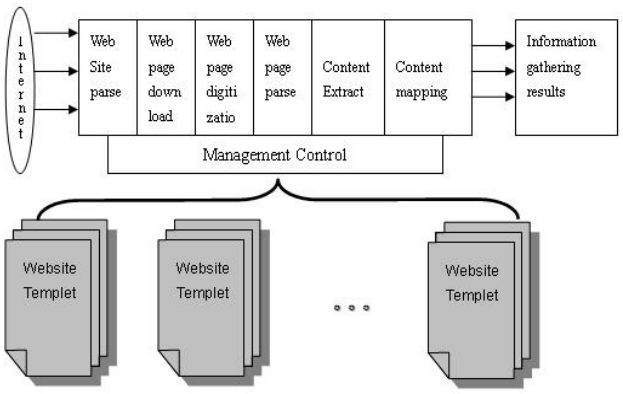

Fig.1 Flow chats of information gathering of WebCPS

The entire gathering process of WebCPS depends on the close cooperation of the gathering engine and the website template. The website template takes charge of the completion of the custom-made function which the user has made, and the gathering engine has the responsibility to read the website stencil, carries on the gather and extract work to the goal website according to the instruction of the website stencil. The descriptions of the six steps processes are as follows:

(1) Website parser

This step completes the analysis to the main body structure of the website, and obtains the website structure characteristic

(2) Webpage downloading

This step completes the integrity downloading of the goal webpage text and the multimedia data.

(3) webpage elements digitization

This step completes the transformation function from the HTML to the XML page, which causes the non- structure HTML data transform to the structure XML data, and digitizes the webpage element according to the inner placed algorithm.

(4) Webpage parser

According to the set of the webpage template, pick up the gathers of the goal element according to its position in the webpage map.

(5) content extracts

Extracts the content information (including the multimedia data) of each element from the goal element gathered, and carries on the on-line processing to the element according to the user's demand to the final information processing.

(6) Content mapping

Classified mapping to the extracted information flows according to the users' demand (or the template setting). The mapping includes two contents. One is to mapping the content which has been extracted to the memory and/or the inquiry classification as the user request according to the natural classification; the other is to mapping the XML document content to the corresponding database table.

It can see from the six steps of the webpage processing flow above, the special filtering and processing model to the webpage data of WebCPS and the common "URL filtering + the full text search" model have the basic difference. The design of the WebCPS is more advantageous to carry on the processing to the complex webpage, and is also advantageous to operate to the data on-line, and reduce the system resources that the trash data has taken up and so on.

The WebCPS system can provide many kinds of gathering service of the agricultural Internet website for the customer, to the website which adopted many kinds of websites technology such as the " Form submit," "JavaScript show", "the entry address changed with the time", "has the Session control"; "ASP, JSP, CGI, PHP and many kinds of Web programming language", "drop-down list accessing entrance" and so on, the system is able to carry on the information gathering, simultaneously it can download the many kinds of the file type which include not only the pictures of multimedia data but also the Word documents, MP3,MPEG and so on.

3 Conclusions

The WebCPS system has initially realized the basic function of the Quzhou agriculture information service website. At present the Quzhou agriculture information service website can automatically gather the important information from dozens agricultural websites in domestic, and the information can be automatically sorted and issued. Besides, the website also can issue its own information. Because of using the information automatically gathering and the issuing system, the workload of the original website which needed 3- 5 persons, can be completed by only one person now. So it has enhanced the working efficiency and carried on the scientific management to the development of the website.

References:

[1] Xiang Jinhui,Xiao Ling, et al. The situation and evaluation of China agriculture information website development, Agriculture Network Information, 2006. Vol.2, pp:4~7

[2] Hu Jinyou,Zhang Jian,You Longyong. Situational Analysis of Agricultural Information

Websites in China, Journal of Agricultural Mechanization Research,2005.Vol.6,pp:38-40
[3] Li Xueyong, Ou Yang Liubo, et al. Comparative Research on Web Spiders' Searching Strategies of Topic Specific Search Engine, Computing Technology and Automation, 2003,Vol.22,No.4 pp: 63~67.
[4] Ou Yang Liubo, Li Xueyong, et al, Survey of Searching Strategies of Web Spiders, Mini-micro Systems , 2005. Vol.26,No.4 pp: 703~706
[5] Zhang Weifeng, Xu Baowen, et al, Overview of the Web Search Engine, Computer Science, 2001. Vol.28,No.9: 24~28
[6] Zou Tao, Wang Jicheng, Zhang Fuyan. Design and Implementation of Information Gathering System Based on WWW, Journal of the china society for scientific and technical information, 1999. Vol.18,No.3: 195~201
[7] Shen Hedan,Pan Yanan,Shao Liangshan, A Study for Search Engine, Computer Technology and Development, 2006,Vol.16,No.4,pp:147 150