Design and Application of Marine Information Retrieval System

Dongning Jia, Yi Chen
Institute of information science and engineering
Ocean University of China
Qingdao, China
e-mail: me@jiadn.com

Yanping Cong*
Institute of information science and engineering
Ocean University of China
Qingdao, China
e-mail: congyp@ouc.edu.cn

***Abstract*—Marine information resource is very important in the development of Marine science field. However, in our country, the marine information resource is lack of effective organization and management. It leads resource disorderly dispersed and makes the marine information using and sharing with low efficiency. Those all caused a huge waste of investment. In this paper, aiming at problems of marine information, we use the web information extraction algorithm to extract marine information in the Internet. This web information extraction algorithm is based on the VIPS algorithm, using the structural and visual features of the web page to extract information. Then based on the metadata theory, we integrate the information and data resources from bottom to top. Finally we store the information in a database and establish a unified and complete marine information retrieval system.**

Keywords-marine information retrieval system, metadata and formal description model, DOM analysis, visual feature

I. INTRODUCTION

With the rapid development of technology, the marine information becomes an international public platform of innovative activities. Our country has always paid attention to marine informatization. However, with the implementation of the "sea power" strategy, the present level of the marine informatization cannot meet the demands. The existing problems and contradictions are increasingly prominent.

Lack of an effective marine information management system, which cause the unclear and inadequate current situation of marine information.

Due to benefits conflict, marine departments or institutions do not share marine information, causing resources scattered and wasted.

Marine information is lack uniform standards and complete service system. It cannot support the comprehensive utilization of marine information and affect the quality and effectiveness of it.

How to regulate the management of marine information and how to make it maximize utility in marine strategy become a crucial problem. Minghua Zhang proposes a multi-source heterogeneous marine data integrated management platform based on B/S framework and the thought of delamination [1]. Feng Zhang use the concept of directory service, according to the features of marine information shared, to study the directory service of marine information based on metadata [2]. Huifen Xue introduce several international marine metadata and make conclusion, analysis and comparison of them [4].

Thus, we propose an information extraction algorithm to extract the marine information in the Internet. Meanwhile we propose a formal description model based on metadata to describe and integrate all marine information in the Internet. Finally we can store it in a database and establish a unified and complete marine information retrieval system.

II. PLATFORM OVERVIEW

This system mainly includes two parts:

- Aiming at marine information in the Internet, we propose a formalized description and organization model based on metadata in order to describe, organize and manage marine information.
- We propose a web information extraction algorithm based on VIPS to extract marine information in the Internet. Then we integrate the information based on the model to form a unified standard format and store it in the database. Finally we develop a marine information retrieval system.

A. Information formal description model

We learn the concept of metadata and apply it to the description and organization of the marine information in the Internet. We use the set to describe the web information and analyze the structure and content of web page. Finally we determine all elements of the set. After formal description and organization, the structure of information becomes clear. It has some advantages such as consistency, simplicity, computer understandability, extensibility and sharing.

TABLE I. TABLE 1 FORMAL DESCRIPTION MODEL OF MARINE INFORMATION

```
W = { F,     //summary of the web information
      A,     //attribute of the web information
      M,     //subject of the web information          }
F = {  keywords,
       description,
       title                                           }
A = {  author,
       date,      //information published time
       classification                                  }
M = {  article_title,
       article_content,
       article_img     //picture of information      }
```

In the end, according to the description model, we design the database and store the extracted marine information.

B. Structure of marine information retrieval system

Marine information retrieval system is mainly designed and developed based on the database of marine information. First we use web information extraction algorithm to extract marine information, then filter and integrate it to form the standard database. Retrieval function mainly uses the description model to search the information, such as keywords, time, author, classification and so on.

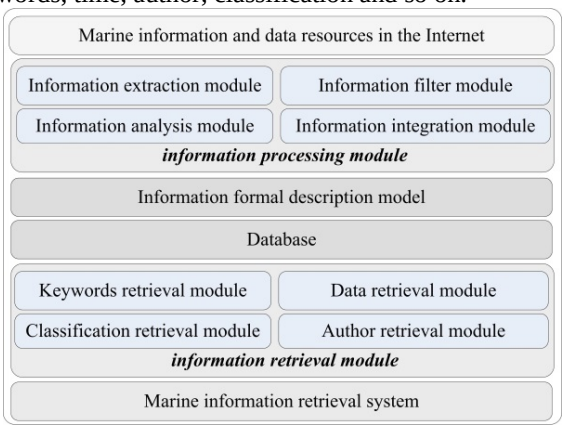

Fig. 1 Structure of the system

This system is based on all marine information in the Internet. The information source is abundant and we solve the scattered and low shared problems. Finally we can establish the unified standard database and develop the marine information retrieval system, providing information and data support for research of marine field.

III. CRITICAL TECHNOLOGY

According to the demand of this system, the design and implementation of it mainly include two critical parts:

- We propose an information extraction algorithm in view of the structural and visual features of the web page based on VIPS algorithm. The web page is chunked by structural and visual features. Then we extract the structural and visual features of each block to form block features. We use it to filter and extract the information.
- We integrate the blocks which have similar features based on marine metadata and formal information description model. We can form a unified standard marine information and store it in a database.

A. Design of web information extraction algorithm

This algorithm is based on VIPS [6]. We select the view of structural and visual features to analyze the web page and extract the web information. According to the structural features of web page and the visual perception of human eyes, it shows the importance and position of the information. We can recognize and process it through web page analysis using computer algorithm. Then we can distinguish between important information and noise information. This algorithm is mainly divided into three perspectives:

- Analysis based on DOM tree.
- Analysis based on HTML tags.
- Analysis based on visual features of page and text.

1) Analysis based on DOM tree

We use the DOM tree as the background technology to analyze the web page. DOM tree reflects the level structure of the web page and it is very easy to be operated. We can preliminary process the web page to be chunked into blocks based on or <table> nodes as unit. Then we can simplify the DOM tree structure of the web page. Finally we can recognize and process the blocks by other features.

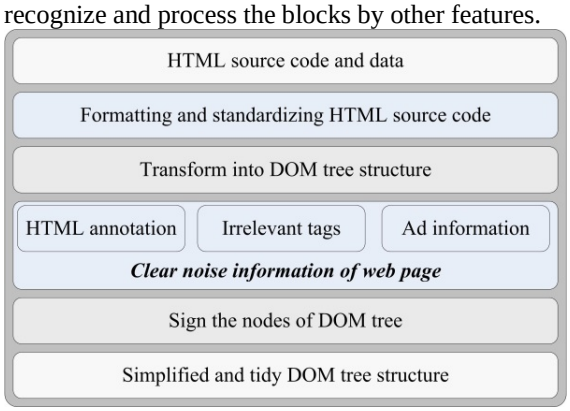

Fig. 2 processing of DOM tree

2) Analysis based on HTML tags

The web page is made up of tags and their mutual nested structure. In the HTML tags, it can be divided into two types according to the different attributes: structural tags and visual tags. Structural tags are related to the page layout. We analyze them mainly includes such as <table>, , <tr>, <td>, <hr> tags. We can divide the web page into different blocks based on them. Visual tags are mainly used to decorate the information and rule the size and position information of blocks, which provide visual information. However because of the web page writing technology progress, visual tags may exist in the web page or in CSS files. We need to extract it from the CSS files if necessary.

We analyze them mainly includes such as <font>, <h1>-<h6>, , <style>, , <p>, tags.

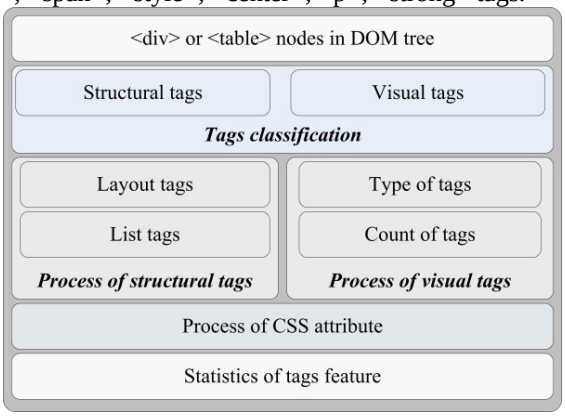

Fig. 3 processing of tags

3) Analysis based on visual features

The visual features of the web page not only expressed by HTML tags, but also expressed by features of text. Both of them have a big effect on the visual expression of web page and we can extract and analyze the visual expression. It is very important for recognizing and extracting information and it can improve the accuracy of the extraction information.

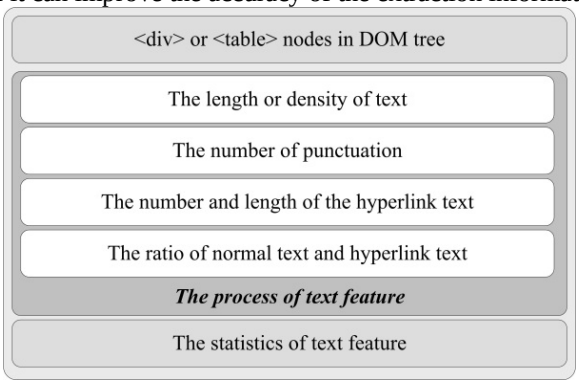

Fig. 4 processing of text visual features

B. Cluster and integration of similar blocks

After processing of web page, the information is divided into many blocks based on <table> or . Meanwhile, we extract and integrate the tags and text features of the blocks, then we can eliminate and prune the noise information. When clustering the similar blocks, we improve the VIPS algorithm according to our demands. We use or <table> nodes as unit of information blocks and integrate them from bottom to top according to the all features of information blocks and pre-determined threshold. Finally we integrate the similar blocks according to the block features and information formal description model.

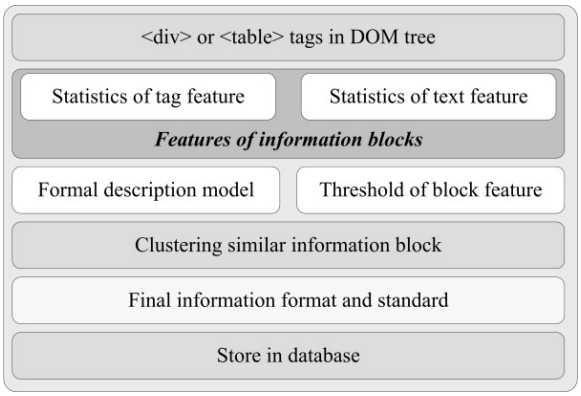

Fig. 5 processing of clustering similar block

In the process of information feature extraction, we need to use mathematical method to describe it. So after statistics of block features, we need to eliminate noise information and cluster similar blocks according to it. This is related to the block threshold. The selection of the threshold has an important effect on the accuracy of extraction results. How to choose a reasonable threshold is the key of our algorithm. We solve it using threshold self-learning correction methods.

1) Threshold of tags feature

We randomly select a number of web pages and record the type and number of tags in each web page. The type of tags is manually determined, including all kinds of visual tags. Then we get the minimum value and average value of each type.

Finally we determine the threshold self-learning correction formula of tags (1):

$$a_n = (\min * n) \quad (1)$$

a_n is threshold; min is minimum value; 1 <= n <= average value.

2) Threshold of text feature

After extracting the HTML code, we clear up all HTML tags and record the number of text. Meanwhile we also extract the hyperlink text and record the number of them.

Finally we determine the threshold self-learning correction formula of normal text (2):

$$b_n = (n * \text{TextAll} / (n + 1)) \quad (2)$$

b_n is the number of normal text; TextAll is the total number of normal text; 1<= n <= 10.

The threshold self-learning correction formula of hyperlink text (3):

$$c_n = (\text{HtextAll} / (n + 1)) \quad (3)$$

c_n is the number of hyperlink text; HtextAll is the total number of hyperlink text; 1<= n <= HtextAll.

IV. System Evaluation

The marine information retrieval system is mainly developed based on marine information in the Internet. We

use the information extraction algorithm to extract and integrate it. Finally we establish a unified and complete database, providing data support for marine areas. The advantages of this system mainly have the following several aspects:

- Completeness. This system strives to extract and integrate all marine information in the Internet, forming a complete marine resources database.
- Synchronicity. This system extracts the marine information from the Internet. It can maintain consistency and synchronicity with resources in the Internet and need not to be manually updated.
- Standard and sharing. After extracting the marine information in the Internet, this system uses the metadata and information formal description model to format and standardize the marine information. We can establish a unified data format and standard, improving the sharing of marine information and data resources.

V. CONCLUSIONS

Aiming at problems of ocean informationization, we propose an extraction algorithm to extract the marine information in the Internet based on VIPS algorithm. Then cooperating with metadata and information formal description model, we organize and integrate the marine information, establishing a unified and complete marine information retrieval system. This system has the advantages of Completeness, Synchronicity, Standard and sharing, improving the usability of marine information and data resources.

ACKNOWLEDGMENT

This work was financially supported by Qingdao innovation and entrepreneurship leading talent project (13-cx-2), Qingdao strategic industry development project (13-4-1-15- HY) and Shandong province science and technology project (2013GHY11519).

REFERENCES

[1] Minghua Zhang, Dongmei Huang, et al. Research and Build of Multi-source Heterogeneous Mass Marine Data Management Platform [J]. Journal of Marine science, 2012, 11:110-115. In Chinese.

[2] Feng Zhang, SiHai Li, et al. A Preliminary Design and Study on Directory Service System of Marine Information Resources [J]. Journal of geographic space information, 2008, 81:81-84. In Chinese.

[3] Kuiying Chen. Speed Up the Pace of Marine Information Construction to Improve Sea Power [J]. Journal of Marine information, 2004, 02:6-8+5. In Chinese

[4] Huifen Xue. Analysis of Several International Marine Metadata [J]. Journal of Marine Information, 2004, 03:25-28. In Chinese.

[5] Shaohua Lin. Development Idea of Marine Information Technology and Service in Our Country [J]. Journal of Marine Information, 2002, 01:8-10. In Chinese.

[6] Cai D, Yu S, Wen J R, et al. VIPS: a Vision-based Page Segmentation Algorithm[R]. Microsoft Technical Report. MSR -TR-2003-79, 2003

[7] Lim S, Ng Y. A Heuristic Approach for Converting HTML Documents to XML Documents [C]. Proceedings of the Sixth International Conference on Rules and Objects. London. England .July 2000. 1182-1196.