Automatic Special type Website Detection Based on Webpage Type Classification

Wang Zhulong¹, Yu Hao², NISHINO Fumihito³

FUJITSU R&D CENTER CO., LTD. Room B1003, Eagle Run Plaza No.26 Xiaoyun Road Chaoyang District Beijing, 10006, China

¹dragon,²yu}@frdc.fujitsu.com; ³nisino@flab.fujitsu.co.jp

Abstract. String-based retrieval techniques sometimes cannot meet requirements of rough retrieval, or requirements for detecting information in a particular domain from the Internet. In this paper, a new website information detection system based on Webpage type classification is proposed for searching information in a particular domain. To filter redundant content of Webpage, an original algorithm of Main Content Block Extraction is also developed. In this algorithm, relationships of WebPages in the same website are taken into account. Experiment results show that our detecting method is effective and realizable. A detection accuracy of 97% can be reached which can satisfy many applications. A detection experiment on Japanese lyric websites was successfully realized by applying our method.

1 Introduction

Currently, Web information processing is mainly focused on the research of information searching and mining technology itself (Gudivada et al.,1997; Brin et al.,1998). Web information searching systems such as Google or Altavista, can help users to find out relevant documents on the Web, while Web data mining, like IBM Intelligent Text Mine, makes it possible to discover implicit patterns on content or structure from large volume of Web documents. However, although most of these systems mentioned above can provide users with some efficient way to obtain resources and knowledge on the Internet, these string-based retrievals and mining methods can not satisfy some special applications, such as the requirements of rough retrieval or special type websites detection from the Internet.

To solve this problem, we propose a new website detecting strategy, which can efficiently detect information involved a particular domain i.e., lyric, news, sports, travel, poesy, products information or games. Here, we view the Webpage type as a kind of page which mainly provides some special domain information. The technology of Webpage processing is the basis of our websites detection. Unfortunately, different websites composers always have their own habits and different websites also have different information to serve users. Most of websites not only have their main content, also include a large volume of redundant information, such as copyright, navigation, multimedia linkages, hyperlinks and so on. So the structures of WebPages are always
complex. This kind of redundant information greatly decreases the system detecting precision. To solve this problem, an original algorithm of Main Content Block Extraction is designed in this paper. Base on this method, a practical application system to detect Japanese lyric website is produced. In our experiment, 3000 WebPages from 60 websites are collected as the Experimental Data. The experimental result shows that both the recognition accuracies of the lyric websites and the non-lyric websites are greatly improved by the algorithm, especially for non-lyric websites, 23.7% improvement on average is obtained. Further analysis shows that the system detecting performance can be improved from 92% to 97%.

Our contributions: We propose a special type website detection method basing on Webpage type classification. In order to improve the recognized accuracy of the WebPages, an original algorithm of Main Content Block Extraction is also developed, which takes the relationships among the WebPages within the same website into account and effectively extract their common templatized information from the WebPages by calculating their block’s contents similarities (Arslan et al., 1999) after parsing the WebPages structures into the DOM trees. A practical application system of Japanese lyric website detection is also produced to validate our website detection methods.

The rest of this paper is organized in the following way: Section 2 introduces the mechanism and the detecting framework of the special type websites detection method. Section 3 is focused on describing the algorithm of Main Content Block Extraction and relevant processing basis. Experiment evaluation and analysis are described in section 4. And the last section presents our conclusion.

2 The Mechanism and Implemental Framework of Special type Website Detection

Different websites always provide customers information in different domains, such as lyric, news, sports, travel, poesy, products information, games and etc. There are a large numbers of websites in the whole internet, with abundant of topics and information. So, how to automatically detect and search out the websites providing some particular domain information becomes an important task. In this paper, the detection of special type website is defined as the process of automatically detecting and searching out the websites providing some special domain information from the Internet. The following is our detecting strategy: Use the Web collector Crawler to collect Webpage from the Internet, as many as possible, and save them into local disks. Then classify the Webpage collected into different groups according to information of their URLs structure. Every Webpage in each group will be estimated, by the method of Multi-Webpage Cooperative Process, to see if it contains any information being interesting to the customers or not, to estimate if the corresponding website is the one the customers want. Here, it is necessary for us to raise the following two points particularly. One point is that: the contents of all Webpage within the same group must contain information of the same domain. The other is the classification of the WebPages in terms of their URLs structure information can be seemed as the pre-classifying on
the recognized samples, which provides foundation and possibility to the following system recognition

In our system, we suggest the detection of the special type website can be carried out should, firstly, classify all Webpage into different groups (one website may be classified into several groups), and then check each group to see if it has information in that particular domain. If so, we regard this website as a website with particular information which the customer interested in and should be included into the target database.

In our methods, we believe that we can increase the accuracy of detection by analyzing the structures of the Webpage in a cooperative processing way. We take structural relations between WebPages within the same group into account, since many websites, especially those, which are professionally designed, always have the same Webpage template and contents. Therefore, an original algorithm of Main Content Extraction (details in section 3), is designed to make the analysis of structure of the Webpage easy. The experiment result suggests that this algorithm can effectively filter the redundant information and greatly improve the system detecting accuracy. The framework for detecting the website information in particular domain is showed as below:

![Webpage Detection Framework](image)

As shows in the Fig 1, to decide the Webpage’s information type, the corresponding information character recognized methods should be introduced to our framework. For example, in our Japanese lyric website detection experiment, for each group, to effectively tell whether it contains lyric information or not, the characters of lyric description, such as its keywords, rhythm, content repeat, and punctuation (Wang et al., 2002), have been taken into account. Finally, all detection results of the WebPages will be recorded into two databases, WebPagecorrect and WebPagemis. And we need to work out the recognition accuracy of each group, and then, get a recognition accuracy threshold, AccuracyThresholdWebPage, by the statistical method to achieve the best system recognizing performance.


3 The Main Content Block Extraction Algorithm

3.1 Content Block: Information Block

Various media information can be embedded into WebPages, including images, videos, text. Webpage may contain many information cells, called Information Elements, which are always formatted by the HTML Tags, to transmit independent communication information.

Fig. 2. Webpage’s Information Blocks

All Information Elements would be organized and organically distributed into the Webpage in a convenient way to users. Considering each information element as an item and a set of information elements sharing the same characters in significance as an item set, then each item set has its peculiar function etc., navigational function and occupies an unattached area in the look and feel. In the paper, these item sets are viewed as Information Blocks. And we define in the following way. Therefore, one Webpage can be seen as made up by different Information Blocks.

Definition 1: [The Information Block] An Information Block is a set of Information Elements which (1) has different function from other Information Blocks in communication transmit; and (2) occupies an independent area of Webpage in the look and feel.

3.2 HTML Parsing Tree

Webpage organized by HTML is a linearization of a document. However, its structure actually looks like a tree (also called as Document Object Models or DOM). In reality, HTML pages are usually composed with a few simple patches, instead of being well
constructed. In this paper, each page is represented by a DOM tree and the investigation will be moved to the corresponding DOM tree of the webpage. In the DOM tree, each HTML Tag represents a non-leaf node and each information element represents a leaf node. After parsing, we can draw out the parsing tree as following.

![Fig. 3. Parsing Tree of the Webpage Showed in the Figure 2](image)

Fig 3 shows that square leaf nodes represent information elements and circle nodes represent HTML Tags, while triangle leaf nodes are the special HTML Tags leaving. From the figure, it is easy to know that two bordered elements in browser may have a far path in its parsing tree, which is helpful for us to divide Webpage into different information block.

### 3.3 Leaf Node Mergence

Comparing Webpage showed by browser to its parsing tree, we can find that information elements among the same information block have the same minimum sub-tree in the parsing tree, and that two information elements in different information blocks have a further distance through the tree paths. In order to find the relations between information blocks through parsing tree, we can pre-process the parsing tree and discard some irrelevant leaf nodes, which are different from the information elements node, like the HTML Tags i.e., `<BR>`, `<HR>`, `<AREA>`. There are two steps to process the parsing tree:

- “N” denotes a node of parsing tree; “DN” denotes a node, which is not an information element but is still a leaf node in the parsing tree; “LN” denotes a node, which is an information element and also a leaf node in the parsing tree.

  Traveling through the parsing tree in Post Order with Depth-First manner and check the node in the following way:
The first step:
   a). If a node N is not a leaf node of parsing tree, do nothing and continue;  
   b). If a node N is not a LN node, then delete this node and continue;  
   c). After this process, all the DN nodes will be discarded.

The second step:
   a). If a node N is a leaf node, do nothing and continue;  
   b). If the parent of node N has only one child and the node N has only one leaf node, then:  
      1). Delete node N;  
      2). Let the sun of node N as one of the sons of his parent node and take its order in the 
         sibling nodes.  
      3). Continue traveling; After discarding the inconsequential nodes, we can get a tidy parsing tree showed as Fig 4.

   If context of all leaf nodes with the same parent can be strung together, and if the 
   leaf nodes strung together is called as information strings, then we will obtain nine 
   information strings. Among these nine information strings, one is Webpage title in-
   formation string, while the remaining eight information strings have the same contents 
   as the eight Information Blocks’.

![Fig. 4. Parsing Tree of the WebPage Post-process](image)

3.4 String Similarity Calculation

An efficient technique for measuring string similarity is to find the edit distance be-
 tween two strings, which can be calculated by applying a dynamic programming algo-
 rithm (Arslan et al., 1999). Edit distance is defined as the minimum cost of transform-
 ing strings X and Y of lengths m, and n respectively with m>n, through a sequence of 
 edit operations which are of three types: insertions, deletions, and substitutions of 
 symbols. The cost of each operation is correspond to a given weight function, which 
 can be defined basing on the effective application. The implementation in this paper, a 
 normalized edit distance algorithm(Arslan et al., 1999), which brings a amortized 
 weight (Arslan et al., 1999) for the given edit sequence and has an O(mn log n) worst-
case for uniform weights to operation function, is applied. Experimental results sug-
gest that for our application normalized edit distance is better as a similarity measure
than ordinary edit distance, which ignores the length of the sequences.

3.5 Templatized Information Block and Main Content Block

A template is defined by Yossef and Rajagopalan as: template is a pre-prepared mas-
ter HTML shell page that is used as a basis for composing new WebPages. The con-
tent of the new pages is plugged into the template shell, resulting in a collection of
pages that share a common look and feel (Yossef et al., 2002). On the other hand, a
templatized webpage can be seemed as a webpage making up by some Information
Blocks, which shares the common information and frame. A integrated webpage based
on templatized webpage is formed by templatized webpage’s Information blocks,
called Templatized Information Block, and the main information plugged into later,
called Main Content Block. To users, the Main Content Block provides dominating
information, while templatized Information Blocks just offer relevant auxiliary pur-
poses and do nothing contribution.

3.6 Templatized Information Block Extraction

Finding the templates of the websites can be viewed as an instance of the frequent
Information Blocks by counting paradigm. The size of the web prevents us from using
a priority to identify template instances. Alternative methods such as the elimination
generation method also fail (Yossef et al., 2002; Agrawal et al., 1994). In this paper,
we propose a new templatized information block extraction algorithm, called multi-
Webpage aiding to template extraction algorithm, which can help us to find the te-
plates by processing multi-Webpage simultaneously. Our task is to extract the com-
mon templatized strings from the Webpage.

We use some pages to extract the templatized Information Blocks. Suppose $n$, a
threshold ($n=3$; in our experiment), is the number of WebPages as study set. And $m$ is
the numbers of Information Blocks of the $n$th Webpage. Then a two dimensions data
structural chain can be established to denote all the information strings of the $n$ Web-
Pages. Among the two dimensions, one dimension is to show the information strings
with the same Webpage and the other is to the number of WebPages, The node $N_{ij}$ ($1
= i = n; 1 = j = m$) denotes the $j$th column node of the $i$th row chain in the two di-
mensioins chain. Then, the main step algorithm can be described as: By scanning through
the chain in the sequence of Left-to-Right and Top-to-Bottom, for each node $N_{ij}$ with-
out being marked, searching a node, which has the maximal value of strings similarity
with $N_{ij}$, within the $i$th row chain ($1 = i = n$ and $i < i$). If the similarity is beyond the
given threshold (0.3 in our experiment), marking a Matched Sign of that node which
will be ignored in next scanning and then putting it into the templatized strings chain.
After scanning through all the two dimensions chain, the templatized strings nodes are
been copied to the templatized strings chain.
3.7 Webpage Main Content Block Extraction

The Main Content Block is the main concerned Block in the extraction. After getting the website’s templatized information strings, it is easy to extract main information string by calculating each two strings’ similarity. If the value of two string’s similarity is larger than the given threshold value, they will be seemed as a matching pair and discarded as it is considered as the templatized information string. The remainder information strings of the Webpage chain is the contents of the Main Content Block.

4 Experiment and Evaluation

Based on the algorithm of the Main Content Block Extraction and with development of the website information detection technique, a practical detection system for Japanese lyric websites is developed. To evaluate its performance, we collected about 3000 WebPages from 60 different websites as the Experimental Data and classified them into different groups in terms of their URLs structure information. We used the symbols, Accuracy_{Lyric-WebPage} and Accuracy_{Non-lyric-WebPage}, to evaluate recognition precision of Webpage in each group. And the performance of the detection system will be worked out with the symbol, Accuracy_{WebSite}, by a given recognition accuracy threshold (AccuracyThreshold_{WebPage}).

All the experimental data mentioned above were checked by human and were divided into two classes: lyric groups and non-lyric groups. Here, as we known that all the Webpage in the same group are coming from the same website and would have the same recognized characters, lyric or non-lyric.

To lyric groups, the Webpage recognition accuracy:

\[
\text{Accuracy}_{\text{Lyric-WebPage}} = \frac{\text{LyricWebPage}_{\text{correct}}}{\text{LyricWebPage}_{\text{correct}} + \text{LyricWebPage}_{\text{miscorrect}}}
\]  

Where, \(\text{LyricWebPage}_{\text{correct}}\) is the number of lyric Webpage being recognized correctly; \(\text{LyricWebPage}_{\text{miscorrect}}\) is the number of lyric Webpage being mistaken as non-lyric Webpage.

To non-lyric groups, the Webpage recognition accuracy:

\[
\text{Accuracy}_{\text{Non-lyric-WebPage}} = \frac{\text{NonLyricWebPage}_{\text{correct}}}{\text{NonLyricWebPage}_{\text{correct}} + \text{NonLyricWebPage}_{\text{miscorrect}}}
\]  

Where, \(\text{NonLyricWebPage}_{\text{correct}}\) is the number of non-lyric Webpage being recognized correctly; \(\text{NonLyricWebPage}_{\text{miscorrect}}\) is the number of non-lyric Webpage being mistaken as lyric Webpage.

The system’s detection accuracy:

\[
\text{Accuracy}_{\text{WebSite}} = \frac{\text{LyricWebSite}_{\text{correct}} + \text{NonLyricWebSite}_{\text{correct}}}{\text{LyricWebSite} + \text{NonLyricWebSite}}
\]

Where, \(\text{LyricWebSite}_{\text{correct}}\) is the number of lyric websites recognized correctly; \(\text{NonLyricWebSite}_{\text{correct}}\) is the number of non-lyric websites recognized correctly;
LyricWebSite is the number of all lyric websites in the test set; NonLyricWebSite is the number of all non-lyric websites in the test set.

Besides, in our experiment, we will use an accuracy threshold to achieve the most excellent system recognition performance.

4.1 The Algorithm Experimental Results

To verify the effectiveness of the Main Content Block Extraction algorithm proposed in the paper, we adopted the symbols $\text{Accuracy}_{\text{Lyric-WebPage}}$ and $\text{Accuracy}_{\text{Non-lyric-WebPage}}$, the ratios of the number of correct recognized Webpage over the number of the total Webpage in each group, to evaluate the recognition results. Within each group, about 50 Webpages from the same website are stored and the corresponding accuracy value is calculated. Simply and intuitively, we divide all the websites into two parts, lyric websites and non-lyric websites by human checked, and draw respectively figures with the accuracy value on the Y-Axis by the ascendant. The results of the recognition accuracy with and without the algorithm of the Main Content Block Extraction (MCBE) are shown in the Fig 5 and Fig 6.

![Fig. 5. The compared results for lyric websites](image-url)
Fig. 6. The compared results for non-lyric website

Each website is signed in above figures by either circles denoted with MCBE algorithm, or asterisks denoted without the MCBE algorithm. Intuitively, we use the curves to connect all the websites with same recognized characters as showed in the figures. From Fig 5 and Fig 6, we can conclude that both recognition accuracies of the lyric websites and the non-lyric websites are greatly improved by the algorithm of the Main Content Block Extraction. Especially for non-lyric websites, 23.7% improvement of the improved accuracies of 30 websites on average is obtained. The curves in these two figures also imply that the algorithm is more effective in recognizing non-lyric websites than lyric websites. However, it is not difficult to illustrate this result, since non-lyric WebPages always contain more complex and divers structures and contents than lyric Webpage.

4.2 Website Recognition Experiment and Evaluation

Performance of the system will be evaluated by the measure of effectiveness, that is, to see how effective the system can provide accurate and creditable recognition result and how useful the system in the other applications is. We will use accuracy defined in formula (3), the ratio of the number of correct recognized websites over the number of the total websites, to evaluate the system’s performance. The symbols drawn in Fig 7 denote the recognition accuracies of all websites by the algorithm of Main Content Block Extraction. From the figure, we can see that both the lyric websites and non-lyric websites are distinctly divided into two parties, although there are three non-lyric websites are still hiding into lyric one. And it is easy for us to distinguish both these two classes by using some classification algorithm to get an accuracy threshold. In Fig 7, a line at 92% is draw out, which been proved can get the best system accuracy. And, in order to show improvement of system detecting performance without and with the MCBE algorithm proposed in this paper, we also draw out two performance curves respectively, under different Threshold in Fig 8.

As shown, the real line denotes the detecting accuracy by using our MCBE algorithm and dashed line show the detecting accuracy without it. From the figure, we can see very clearly that, after using the algorithm, the system detecting performance is greatly improved at any given threshold, and that, for a given threshold value of 92%, the highest detecting accuracy of 97% can be reached. However, on the given threshold value of 78%, the highest performance 92% can only be obtained without the algorithm. It implies that both the algorithm of Main Content Block Extraction developed and the system detecting performance are effective.
5 Conclusion

It is easy to see that the Webpage type classification based detection method can skillfully cover the shortcoming of the string based one and makes rough retrieval on information in a particular domain possible. The Webpage type classification based method, at the same time, carries out the Webpage classification by a different point view on the information expression. And the URLs structure information of Webpage also can be seen as a good character option on the samples pre-classifying.

From our experiment, we can conclude that the algorithm of Main Content Block Extraction proposed in the paper is effective. It can efficiently filter the redundant content of the Webpage and improve the detecting performance, especially for the non-lyric Webpage. Besides, making fully use the relationships among the Webpage within the same website becomes the key for the algorithm. Evidently, the algorithm and the website detection technique proposed in this paper are also fit to detect some other special domain website, such as: sports, travel, poesy or games.
References