2009 Third International Symposium on Intelligent Information Technology Application

DOM-based Web Pages to Determine the Structure of the Similarity Algorithm

Chunying Kang[1]
[1] College of information science and technology,
Heilongjiang University, Harbin,HeiLongJiang,150080,China
Kcy619@yahoo.com.cn

Abstract

Web data is currently mainly in the form of HTML pages, expressed by the HTML language of Web pages through the browser after analysis is only suitable for people to browse, not suitable for data exchange as a way to deal with by a computer. This article will make web page decompound a DOM tree, then from the DOM tree body root node to start , in accordance with the breadth-first traversal order DOM tree, layer by layer comparison DOM node tree, statistics of its changes, and then the sum of all floors of the changes, If less than a certain threshold, it is structurally similar to two pages, otherwise dissimilar. because this algorithm is only concerned about the page structure information without concern for the content of the page, it has a very high operating efficiency, while the algorithm is not limited to a specific web page, with good versatility.

Keywords*： *DOM; Similarity Algorithm;Web

1. Introduction

With the Internet's rapid development, web access to information has become a primary source of information. Web data is currently mainly in the form of HTML pages, because HTML markup language is just tell the browser how to display it the definition of information, and does not contain any semantics, expressed by the HTML language of Web pages through the browser after analysis is only suitable for people to browse, not suitable for data exchange as a way to deal with by a computer. How from this unstructured, does not contain any semantic document automatically by a computer to extract the necessary information, has now become a research direction. Methods are currently used for a single page of information extraction, these methods required for each page are the same analytical processing, efficiency is not high. We note that the procedure in the generated Web pages often use the same page template to format the content of the same type of information display, which features not only the same type of page to show the form completely similar, but the page exists in a highly structured information. We can structure similar pages grouped into one category, in accordance with its definition of the structural characteristics of such information extraction template rules, and thus a large number of web pages generated through the template information collected efficiently solve. The information extraction method the key is to realize how to determine the structural similarity between pages. This article will make web page decompound a DOM tree, then from the DOM tree body root node to start , in accordance with the breadth-first traversal order DOM tree, layer by layer comparison DOM node tree, statistics of its changes, and then the sum of all floors of the changes, If less than a certain threshold, it is structurally similar to two pages, otherwise dissimilar.

2. Related knowledge

2.1 DOM tree

DOM is the Document Object Model Document Object Model abbreviation. According to W3C DOM specification (http://www.w3.org/DOM/), DOM is a browser, platform, language-independent interface, allows users to access pages of other standard components. HTML documents have been translated into DOM trees, HTML documents composed of all nodes in a document tree. The HTML document, each element, attribute, text, and so represents a tree node. In the DOM tree, the document node as a start node, and thus continue to extend branches in this tree until the lowest level until all the text nodes. Figure 1 express an HTML document tree.

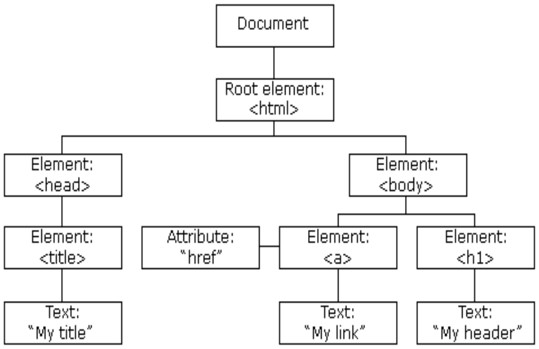


Figure 1. HTML document tree

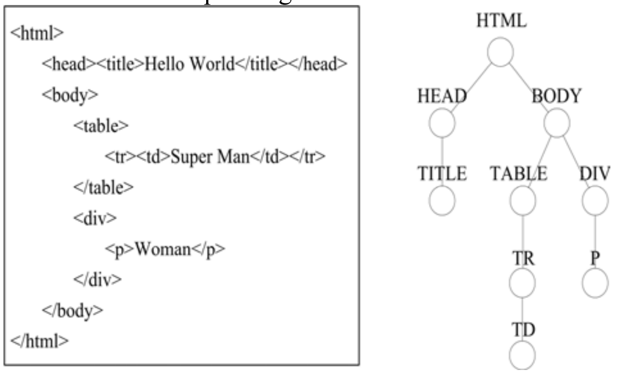
2.2 Webbrowser control

978-0-7695-3859-4/09 $26.00 © 2009 IEEE
DOI 10.1109/IITA.2009.218

The WebBrowser control provides a managed wrapper for the WebBrowser ActiveX control. The managed wrapper lets we display Web pages in your Windows Forms client applications. We can use the WebBrowser control to duplicate Internet Explorer Web browsing functionality in your application or you can disable default Internet Explorer functionality and use the control as a simple HTML document viewer. We can also use the control to add DHTML-based user interface elements to your form and hide the fact that they are hosted in the WebBrowser control. This approach lets we seamlessly combine Web controls with Windows Forms controls in a single application.

3. Algorithm description

According to HTML tags nested relations, a web page in terms of logic can be expressed as a tree. Tag tree nodes express the relationship between the nested. As shown in Figure 2, the map section of the list of HTML code and its corresponding DOM tree.



```
<html>
    <head><title>Hello World</title></head>
    <body>
        <table>
            <tr><td>Super Man</td></tr>
        </table>
        <div>
            <p>Woman</p>
        </div>
    </body>
</html>
```

Figure 2 DOM tree

Determine the page structure based on a similar purpose, we have to deal with the structure of web pages play a decisive role in the layer, so from the beginning of BODY node, because of the contribution of each different node from the beginning of BODY, the higher the level of the tree the greater its contribution to . In the algorithm, we traverse the largest low-rise and on each floor to limit the right to redo the structure is similar to the impact of the larger tag, including TABLE, TR, TD, DIV, SPAN, multiplied by an amplification factor , highlighting its role.

Definition 1 (weight W): page DOM tree on each floor of the contribution of the page similarity measurement, the formula is: $W_{i=}\frac{D-Li}{D}$,i $\in$[1，D].

D: to traverse the tree's largest low-rise

L:tree of current rise

Definition 2(amplification factor α):Tag on similarity measurement, $\alpha_{i,j} \in \{0,1,2\}$,i for the number of layers,j for the position of layer in the I, $i \in [1, D], j \in [1, T_i], T$ for the total number of elements on each layer.When the tag is TABLE、TR、TD、DIV、SPAN， α =2;when the tag is other changed of the HTML tags, α =1;when the tag is no changed HTML tags, α =0.

Definition 3 (change in percentage of P): page DOM tree changes in all levels of measurement, which reflects the structure of web page changes, the formula is:

$$P_i = \frac{\sum(E_k * \alpha_{i,j})}{\sum(E_j * \alpha_{i,j})} * \frac{1}{D} * W_i, i \in [1, D], j \in [1, T_i], k \in [1, C_i]$$

,E express each floor has changed the tag, C express the total number of floors to change the tag.

Definition 4 (similarity S): be used to measure the page structure is similar to the situation, the formula is: $S = 1 - \sum_{i=1}^{D} P_i$, If it is greater than the similarity threshold (referred to as Q), is similar to two pages, otherwise dissimilar.

The algorithm used in the two to be determined parameters D and Q, in order to select the appropriate value, we conducted the following experiment. We have eight different search engine sites, select the 100 search results page similar comparison, by taking different values D, the following table:

Table 1. Cross Reference Table

Web Q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
www.baidu.com	1	1	1	1	1	0.99	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
www.yahoo.com.cn	1	1	1	0.99	0.96	0.95	0.94	0.92	0.91	0.91	0.9	0.89	0.88	0.87	0.85
www.google.com	1	0.92	0.9	0.89	0.88	0.85	0.84	0.82	0.8	0.79	0.77	0.77	0.76	0.77	0.77
www.live.com	0.75	0.88	0.92	0.94	0.95	0.95	0.95	0.94	0.94	0.93	0.92	0.91	0.9	0.89	0.88
www.xinhuanet.com	1	1	1	1	0.99	0.97	0.95	0.93	0.92	0.92	0.91	0.91	0.91	0.91	0.91
www.alibaba.com	1	1	1	1	1	0.99	0.99	0.99	0.98	0.97	0.95	0.94	0.93	0.93	0.92
www.youdao.com	1	1	1	1	1	1	0.98	0.96	0.94	0.92	0.91	0.89	0.88	0.87	0.86
www.youku.com	1	1	1	1	0.98	0.96	0.95	0.94	0.93	0.93	0.93	0.93	0.93	0.94	0.94

Based on Table 1 data we draw polyline Figure 3, we can see that when the ergodic layer is too small (D <7), pages difference in structure has not been fully reflected, and when the ergodic layer is too more (D> 10), the overall structure of the web pages of non-decisive role of the nodes will be too much involved in the calculation, affecting the ultimate value of similarity. According to the experimental results, we take D = 8, Q = 0.8.

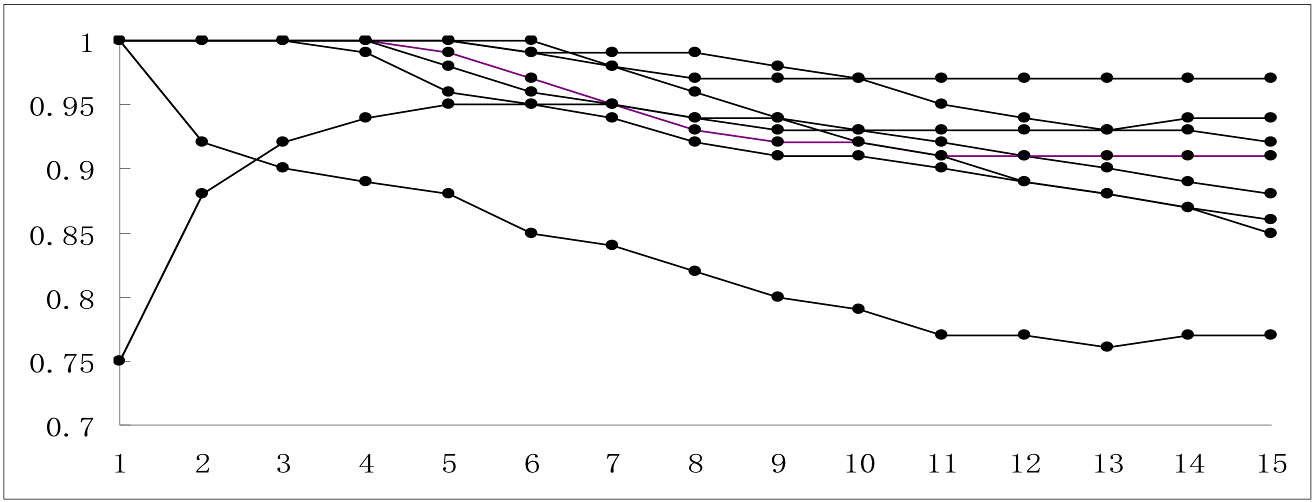


Figure 3. Line chart

4. Algorithm implementation

This article uses Microsoft's Visual Basic as a tool for implementation, using webbrowser control to the target site collection of HTML and convert it to DOM tree, the main code is as follows:

```
'A and B are an array of variables that used to store Web site and the target DOM tree.
'A_T and B_T are used to store Web site and the target DOM tree of the total number of array elements.
'To determine whether the elements are important structural elements.
Private Function SpecialTag(T)
     If T = "TABLE" Or T = "TR" Or T = "TD" Or T = "DIV" Or T = "SPAN" Then
         SpecialTag = True
       Else
         SpecialTag = False
       End If
End Function
'Calculation of similarity Q
Function Q(D)
    Dim P, T, W
    P = 0
    For i = 1 To D
        T = A_T(i)
        If B_T(i) > T Then
            T = B_T(i)
        End If
        Ek = 0
        Ej = 0
        For j = 1 To T
          If A(i, j) <> B(i, j) Then
         If SpecialTag(A(i, j)) Or SpecialTag(B(i, j)) Then
                Ek = Ek + 2
            Else
                Ek = Ek + 1
            End If
          End If
         If SpecialTag(A(i, j)) Or SpecialTag(B(i, j)) Then
            Ej = Ej + 2
          Else
            Ej = Ej + 1
          End If
        Next
        If T > 0 Then
          W = (D - i + 1) / D
          P = P + ((Ek / Ej) / D) * W
        End If
    Next
    Q = 1 - P
End Function
```

5. Conclusion

In this paper, based on the DOM of the page to determine the structure of two similar algorithms, because this algorithm is only concerned about the page structure information without concern for the content of the page, it has a very high operating efficiency, while the algorithm is not limited to a specific web page, with good versatility. However, because of the procedures required to run the course web page DOM tree stored in memory, if the site's HTML too much, it would be more expensive system memory. In further studies, we will take advantage of depositors switching technology inside and outside of this algorithm to reduce the algorithm running on the system environment, hardware requirements, as well as improve the operating speed algorithm.

References

[1] Yang, Y D, Zhang H J HTML Page Analysis Based on Visual Cues[C] // Proceedings of the Sixth International Conference on Document Analysis and Recognition Washington: IEEE Computer Society, 2001: 10-13

[2] Lin Shian hua,Ho Jan wing Discovering informative content blocks from Web documents [C] // Proceeding of the 8th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining Edmonton ACM Press, 2002: 588-593

[3] World Wide Web Consortium. Document Object Model Activity Statement [S/OL].(2007)-[2007-05-12].http://www.w3.org/DOM/Activit

[4] Andy Clark. CyberNeko HTML Parser [EB/OL].(2005)[2007-05-12].http://peo-ple.apache.org/andyc/neko/doc/html/index.html