Mining Structured Objects (Data Records) Based on Maximum Region Detection by Text Content Comparison From Website

G.M. Atiqur Rahaman¹, Md. Mobarak Hossain², Md. Ahsan Arif³, Eyrin Chowdhury⁴, Sharmi Debnath⁵
Computer Science and Engineering Discipline, Khulna University¹, ², ³, Khulna 9208 Asian University of Bangladesh², ³, Dhaka-1230, Bangladesh
atiq99@hotmail.com, mobarak10jan@yahoo.com, mdahsanarif@yahoo.com, anne.toon@yahoo.com, sharmi_ku_04@yahoo.com

Abstract-- At present, a great amount of information on the Web is presented in regularly structured objects. These are known as data records. A list of such objects in a Web page often describes a list of similar items; such as, a list of products to provide their value-added services. Therefore, it has become increasingly necessary to develop an effective process for extracting information from them. In this paper, we present a more effective method to perform the task. The proposed method is able to mine data records not only from a single Web page but also from an entire Web site. The performance of the proposed method is evaluated in respect to the previous methods in the literature. Our experimental results show that the proposed technique outperforms the existing techniques eventually. Finally, we compare the results of the experiments and discuss about the performance of the proposed method in the case of mining structured objects (data records).

Index Term-- mining, Crawling, Web information integration, Web data regions, Web data records.

I. INTRODUCTION

With the tremendous growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools for finding the desired information resources, for tracking and analyzing their usages patterns. These factors give rise to the necessity of creating intelligent systems that can effectively mine knowledge. Data mining from the Web can be broadly defined as the discovery analysis of useful information from the World Wide Web. This describes the automatic search of information resources available on-line, i.e., Web content mining.

The Web is perhaps the single largest data source in the world. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues. Much of the Web information is structured due to the nested structure of HTML code, which we call data records. Mining structured objects or data records from a Web site able to mine all the data records, where data record allows us to integrate information from multiple sources to provide value-added services. In [1] Liu et al. presented a method of Mining Data Records in a web page (MDR). In their methods, it basically exploits the regularities in the HTML tag structure directly. It is often very difficult to derive accurate wrappers entirely based on HTML tags. The MDR algorithm makes use of the HTML tag tree of the web page to extract data records from the page. However, erroneous tags in the HTML source pages may result in building of incorrect trees, which in turn makes it impossible to extract data records correctly. MDR has several other limitations, which will be discussed in the later half of this paper. In [2] Zhai et al. presented a method of Web Data Extraction Based on Partial Tree Alignment (DEPTA) uses visual information to build the HTML tag tree to eliminate the problems caused by HTML tag structure irregularities. It also uses a tree edit distance for identifying similar data records.

In [4] Benchalli et al. presented a method of Visual Structure based Analysis of web Pages (VSAP). In their methods, they have mined data region by using the visual structure of the Web pages. It can be analyzed that, the relevant data region seems to occupy the major central portion of the web page. Inspired from the above papers, we present a new method for mining structured objects (data records) from Web pages. Our method is called MDRMTA (Mining Data Record Based on Maximum Text Analysis). We choose the data region with the help of maximum amount of text content comparison.

II. EXISTING METHOD FOR MINING STRUCTURED OBJECTS (DATA RECORDS)

Extracting the regularly structured data records from web pages is an important problem. Several attempts have been made for dealing with the problem so far. Some techniques have poor accuracy and some techniques do not able to mine non-contiguous structured objects. In this paper, a new method of mining structured objects (Data records) has been proposed future development. The method has been implemented and tested for some sample web pages from various commercial websites. Though a number of researches have been conducted regarding mining data records, but in the specific field of Data Mining, the number of available researches is not enough presently. The following sections describe the most recent research developments on this aspect.

Mining Data Records in web page (MDR) was proposed by Liu et al. in 2004 [1]. In their methods, it automatically mines all data records formed by table and form related tags i.e.,
<TABLE>, <FORM>, <TR>, <TD>, etc. assuming that a large majority of web data records are formed by them. The algorithm is based on two observations. A group of data records are always presented in a contiguous region of the web page, which is formatted using similar HTML tags. Such a region is called a Data Region. The nested structure of the HTML tags in a web page usually forms a tag tree where a set of similar data records are formed by some child sub-trees of the same parent node. In this case, at first HTML tag tree is built by following the nested blocks of the HTML tags in the web page. Then we identify the data regions by finding the existence of multiple similar generalized nodes of a tag node. Finally, we mine the data records from each generalized node in a data region. A problem with the assumption is that, the relevant information of a data record is contained in a contiguous segment of HTML code, which is not always true. So, this method cannot provide high quality results. It also performs poor because the algorithm makes use of the HTML tag tree of the web page to extract data records from the page. However, an incorrect tag tree may be constructed due to the misuse of HTML tags, which in turn makes it impossible to extract data records correctly. Moreover, the proposed method uses additional keyword search, such as ($) or similarity threshold value in order to identify similar data records.

Web Data Extraction Based on Partial Tree Alignment (DEPTA) was proposed by Zhai et al. in 2006 [2]. They presented a new method of two-step strategies. At first we identify individual data records in a page then aligning and extracting data items from the identified data records. Benefits of this method are visual information is used to build the HTML tag tree, to eliminate the problems caused by HTML tag structure irregularities. The method also uses visual gaps between data records to deal with the wrong combinations of sub-trees problem, which occurs in previous method Mining Data Records in the web page (MDR) that was proposed by Liu et al. in 2004 [1]. In case of Partial tree alignment the tag trees of all data records in each data region are aligned using their partial alignment method, which is based on tree matching techniques. This method of constructing a tag tree has the limitation that the tag tree can be built correctly only as long as the browser is able to render the page correctly. This method performs poor in case of computation time, where constructing the tag tree, traversing the whole tag tree and string comparison are available also on the overhead.

Visual Structure based on Analysis of Web Pages (VSAP) was proposed by Benchalli et al. in 2005 [4]. In their methods, it finds the data regions formed by all types of tags using visual cues. The visual information enables the system to identify gaps that separates data records, which helps to segment data records correctly because the gap within a data record of any is typically smaller than that of between data records. By the visual structure analysis of the web pages, it can be analyzed that the relevant data region seems to occupy the major central portion of the web page. The method is based on three observation; A group of data records that contains descriptions of a set of similar objects, which is typically presented in a data region of a web page. The area covered by a rectangular shape that bounds the data region, which is more than the area covered by rectangle bounding other regions, e.g. advertisements and links. The height of an irrelevant data record within a collection of data records is less than the average height of relevant data records within that region. In the first step of this method, the co-ordinates of all the bounding rectangles in the web page are determined. This approach uses the MSHTML parsing and rendering engine of Microsoft Internet Explorer 6.0. This parsing and rendering engine of the web browser deliver the co-ordinates of a bounding rectangle. The second step of the method is to identify the data region of the web page where the data region is the most relevant portion of a web page that contains a list of data records.

III. MINING STRUCTURED OBJECTS (DATA RECORDS) FROM WEBSITE

In this section, we present a new method for structured object mining based on maximum text content comparison. It uses crawling algorithm for extracting all the URLs of a Website. It can reduce the amount of time to mine structured objects. The main idea behind this proposed method is to mine structured objects (data records) from a given address of a Website.

The mining procedure will be performed using the following steps:
- **Step 1:** Crawl the Website of given input URL.
- **Step 2:** Identify the largest text area that contains the maximum amount of information.
- **Step 3:** Mine data region from the largest text area.
- **Step 4:** Mine data records from the data region.

The flow chart of mining data records from a website process has been depicted in the following Fig. 1.

![Flow chart of mining data records from a website](image)

**Fig. 1. Flow chart of mining data records from a website**
A. Crawling a Website

Our proposed crawling algorithm is separated into two parts: crawling with no knowledge base and crawling with knowledge bases. Crawling with no knowledge base is used in the first crawling since the crawler has not had any knowledge yet. Crawling with knowledge base is used in the consecutive crawling. These knowledge bases are progressively built from the previous crawling.

Procedure Crawling (without knowledge base)
INPUT the first URL
Download the page of the selected URL
Read the page
IF this is an HTML containing page THEN
Extract anchor tags
Enqueue the tags in a queue A
IF any tag matches with the input URL THEN
REPEAT
Enqueue the tag in queue B
Check the next tag from queue A
UNTIL no more tags in queue A
ELSE Discard the tag
Procedure Crawling (with knowledge base)
REPEAT
Select anchor tag from queue B
Load and read the page
IF this is an HTML containing page THEN
Extract all the anchor tags of that page
Enqueue all the tags in queue C
IF any tag fully match with the tag of queue B and input URL THEN
Discard it
IF mismatch with the tags of queue B and input URL THEN
Insert the anchor tag in queue C
Check the next tag from queue B
UNTIL no more tags in queue B

One must begin the crawling process using the first pseudo code illustrated in Procedure Crawling (without knowledge base). We assign a Web site address to the crawler, and then it catches all the URL’s of that Web site. The pseudo code illustrated in Procedure Crawling (with knowledge base) takes the each output URL of first pseudo code as input and further fetch another URL of that site which is not get by first pseudo code.

B. Identify largest text area

In every HTML page each and every HTML tag has an associated rectangular area on the screen. This rectangular area is called the bounding rectangle. The bounding rectangle that contains the maximum amount of text is called the largest text area.

Procedure Find maxRect (BODY)
max_rect_list=NULL.
max_rect=body of the total HTML document as a node.
REPEAT
For each child of body tag
BEGIN
max_rect = first child of body tag
Node = first child of body tag
REPEAT
count_max = 0
count_node = 0
count_max = innertext length of max_rect
count_node = innertext length of node
If (count_max < count_node) then
max_rect = node
Endif
Node = next sibling of node
UNTIL (node does not become null)
max_rect_list = max_rect
UNTIL (max_rect does not become null)

C. Mine Data Region

Mining data region using our proposed technique has the following observations.
(a) A group of data records which contains descriptions of a set of similar objects is typically presented in a contiguous region of a page.
(b) The area covered by a rectangle that bounds the data region is more than the half of the area covered by rectangles bounding other regions such as: advertisements and links.

Procedure Find dataRegion (max_rect)
data_region = max_rect_list [0]
For each node of max_rect_list
BEGIN
If (half of innertext length of data_region < innertext length of node) then
  data_region = node
End if
END

D. Mine data records from data region

The output of mining data regions is considered as input of this step.

Through filtering the child of a data region, data record is obtained. In the same time, the outer text length and the no of child of data region is computed. If the outertext length of each node is greater than or equal to average outertext length then keep them otherwise remove them.

Procedure
Find dataRecord (data_region)
Total outertext length = 0
For each node of data_region
  Total outertext length += outer text length of each node of data_region
End
Average outertext length = total outertext length / no of node of data_region
For each node of data_region
BEGIN
If (outertext length of each node > average outertext length) then
  data_record = node
END if
END

IV. EXPERIMENTAL RESULTS AND COMPARISON

In this section, we evaluate the proposed technique. We also compare it with two state-of-the-art existing systems, Zhai et al DEPTA [2] (which is an improvement of Liu et al MDR [1]) and Liu et al MDR [1]. We do not compare it with the method Chang et al IEPAD [5] and the method Crescenzi et al “Towards Automatic Data Extraction from Large Web Sites”[6], as it is shown in Liu et al MDR [1] that Liu et al MDR [1] is already more effective than them. The evaluation consists of four aspects as discussed in the following:

A. Experimental Results

In running Liu et al MDR [1], we used their default settings. Liu et al MDR [1] system was downloaded at http://www.cs.uic.edu/~liub/MDR/ MDR-download.htm. We use the recall measures (which is widely used to evaluate information retrieval systems) to evaluate the performance of our system for mining data records. Recall is defined below:

Recall = CER / TNR

Where, CER is the total number of correctly extracted records, TNR is the total number of records on the page. Recall defines the correctness of the data records identified from the web page. The experimental results are given in Table 1. Correct results (Cor.) are relevant data records present on the page that have been correctly identified. The results obtained after running both Liu et al MDR [1] and MDRMTA are shown in Table 1. The successes of both algorithms are compared in terms of recall.

TABLE I
PERFORMANCE COMPARISON OF MDR AND MDRMTA

<table>
<thead>
<tr>
<th>URL</th>
<th>Total no of data records in a page</th>
<th>MDRI</th>
<th>MDRMTA</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.fagdirect.com/searchelectrosmis.html">http://www.fagdirect.com/searchelectrosmis.html</a></td>
<td>25</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td><a href="http://www.amazon.com/067145937x/reviews">http://www.amazon.com/067145937x/reviews</a></td>
<td>13</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td><a href="http://www.coolLu.com/booceris/coolLu.asp">http://www.coolLu.com/booceris/coolLu.asp</a></td>
<td>15</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td><a href="http://makers.shop.dinar.com/dinar/reis">http://makers.shop.dinar.com/dinar/reis</a></td>
<td>07</td>
<td>07</td>
<td>07</td>
</tr>
<tr>
<td><a href="http://www.pamiell.com/">http://www.pamiell.com/</a></td>
<td>15</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td><a href="http://video.barnesandnoble.com/animation/animation">http://video.barnesandnoble.com/animation/animation</a>...</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td><a href="http://jewelry-watches.pricegrabber.com/wedding-accessories">http://jewelry-watches.pricegrabber.com/wedding-accessories</a></td>
<td>24</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td><a href="http://www.m">http://www.m</a> admits.com/oc- Musical_instruments...</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td><a href="http://international.drostore.com/Default.asp">http://international.drostore.com/Default.asp</a></td>
<td>6</td>
<td>Error</td>
<td>16</td>
</tr>
<tr>
<td><a href="http://www.basicsol.com/sol/aux-emp">http://www.basicsol.com/sol/aux-emp</a></td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><a href="http://www.target.com/pg/browse.html">http://www.target.com/pg/browse.html</a></td>
<td>9</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><a href="http://www.target.com/pg/browse.html">http://www.target.com/pg/browse.html</a></td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>282</td>
<td>178</td>
<td>256</td>
</tr>
<tr>
<td>Recall</td>
<td>63.12%</td>
<td>83.60%</td>
<td></td>
</tr>
</tbody>
</table>
B. Data Region Mining

We compare the first step of Liu et al MDR [1] with our system for identifying the data regions. Both Liu et al MDR [1] and Zhai et al DEPTA [2], are dependent on certain tags like <TABLE>, <TBODY>, etc. for identifying the data region. But a data region need not be always contained only within specific tags like <TABLE>, <TBODY>, etc. A data region may also be contained within tags other than table-related tags like <P>, <LI>, <FORMS>, <DIV> etc. In our proposed system, the data region identification is independent of specific tags and forms. Unlike Liu et al MDR [1], where an incorrect tag tree may be constructed due to the misuse of HTML tags, there is no such possibility of erroneous tag tree construction in case our proposed technique, because the hierarchy of tags is constructed based on the maximum data or information container on the web page. In case of Liu et al MDR [1] and Zhai et al DEPTA [2], the entire tag tree needs to be scanned in order to mine data regions, but MDRMTA doesn’t scan the entire tag tree, but it only scans the largest child node of the <BODY> tag. Hence, this method proves very efficient in improving the time complexity compared to other contemporary algorithms.

C. Data Record Minings

We compare the record extraction step of Liu et al MDR [1] with MDRMTA. Liu et al MDR [1] identifies the data records based on keyword search (such as "$\$"). But MDRMTA is purely dependent on the maximum information container of the web page only. This proves to be very advantageous as it overcomes the additional overhead of performing keyword search on the web page. Liu et al MDR [1], not only identifies the relevant data region containing the search result records but also extracts records from all the other sections of the page, such as advertisement records also, which are irrelevant. Liu et al MDR [1] also sets a similarity threshold value of 0.3 as the default value for the system to identify similar data records, which has been set based upon a number of training pages. Hence, it doesn’t scale well with all web pages. In Liu et al MDR [1], comparison of generalized nodes is based on string comparison using normalized edit distance method. In Zhai et al DEPTA [2], comparison of sub-trees is made by simple tree matching algorithm. However, both these methods are slow and inefficient as compared to MDRMTA where the comparison is purely based on nominal, since we are comparing between largest rectangles of a Web page for mining data record.

D. Time Comparisons

MDRMTA is less time consuming than previous (Liu et al MDR [1], Zhai et al DEPTA [2]) methods. In case of time comparison, we have considered second (sec) as the unit of time. The following table represents the comparison according to time between our proposed method and Liu et al MDR [1]. From the following table we have seen that in some cases Liu et al MDR [1] is less time consuming than MDRMTA. But it is also seen that in those cases MDRMTA get more data records than Liu et al MDR [1]. Again, from the following table it can be seen that, in only two cases MDRMTA can mine less data records than Liu et al MDR [1]. Perhaps the only reason is that, in our method we have considered that, data record is contained by the largest rectangle that holds maximum amount of text. Where, there are some Websites in which data record may contain not only in the largest rectangle but also in other rectangle which is adjacent to this largest rectangle. This comparison of time duration for each URL leads to the conclusion that overall MDRMTA is a time saving algorithm which is able to mine comparatively more data record than Liu et al MDR [1].

Table 2: Time Comparison of MDR and MDRMTA

<table>
<thead>
<tr>
<th>URL</th>
<th>Total no of data records in a page</th>
<th>MDR</th>
<th>MDRMTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time</td>
<td>Time</td>
</tr>
<tr>
<td>filter/electronic/index.asp</td>
<td>21</td>
<td>5.55 sec</td>
<td>10.25 sec</td>
</tr>
<tr>
<td><a href="http://www.youtube.com/serious.com/investigation">http://www.youtube.com/serious.com/investigation</a></td>
<td>67</td>
<td>67.47 sec</td>
<td>67.28 sec</td>
</tr>
<tr>
<td><a href="http://www.porn.com/">http://www.porn.com/</a></td>
<td>15</td>
<td>2.11 sec</td>
<td>6.37 sec</td>
</tr>
<tr>
<td><a href="http://www.humorandtrivia.com/">http://www.humorandtrivia.com/</a></td>
<td>10</td>
<td>0.21 sec</td>
<td>10.19 sec</td>
</tr>
<tr>
<td><a href="http://www.securitywatch.com/directories/internet.cr">http://www.securitywatch.com/directories/internet.cr</a></td>
<td>34</td>
<td>0.26 sec</td>
<td>22.38 sec</td>
</tr>
<tr>
<td><a href="http://www.searchengineoptimization.net/">http://www.searchengineoptimization.net/</a></td>
<td>13</td>
<td>12.34 sec</td>
<td>12.31 sec</td>
</tr>
<tr>
<td><a href="http://www.google.com/e-commerce.html">http://www.google.com/e-commerce.html</a></td>
<td>24</td>
<td>24.34 sec</td>
<td>24.25 sec</td>
</tr>
<tr>
<td><a href="http://www.topolin.co.uk/">http://www.topolin.co.uk/</a></td>
<td>3</td>
<td>1.17 sec</td>
<td>5.15 sec</td>
</tr>
<tr>
<td><a href="http://www.target.com/cgi-bin/searchlocal">http://www.target.com/cgi-bin/searchlocal</a> ...</td>
<td>9</td>
<td>2.11 sec</td>
<td>1.13 sec</td>
</tr>
<tr>
<td><a href="http://www.target.com/cgi-bin/fox">http://www.target.com/cgi-bin/fox</a> ...</td>
<td>15</td>
<td>8.35 sec</td>
<td>15.88 sec</td>
</tr>
</tbody>
</table>

Total: 282 | 178 | 416 | 256 | 415 |
V. CONCLUSION AND SCOPE OF FUTURE WORK

In this paper, we have proposed an approach to mine structured objects from Web sites. Although the problem has been studied by several researchers, existing techniques are either inaccurate or make many strong assumptions. An effective method MDRMTA is proposed to mine the data records in a web page. It is a pure text content analysis oriented method. We think that our proposed technique (MDRMTA) will be able to correctly identify data records from a Website, irrespective of the type of tag in which it is bound. Other methods would work for only tags of certain types. The number of comparisons done in MDRMTA is significantly lesser than other approaches. Further, the comparisons are made on amount of text, unlike other methods where strings or trees are compared. Thus MDRMTA overcomes the drawbacks of existing tag tree based methods and performs significantly better than these methods.

Our proposed method (MDRMTA) is efficient for those Web pages in which only one container is available and largest amount of information is contained by that container. But there remain many web pages where more than one container sustains essential information. Our proposed method (MDRMTA) does not perform effectively for those Web pages.

Extraction of the data fields from the data records contained in these mined data regions will be considered in the future work. The extracted data can be put in some suitable format and eventually stored back into a relational database. Thus, data extracted from each web-page can then be integrated into a single collection. This collection of data can be further used for various Knowledge Discovery Applications such as making a comparative study of products from various companies, smart shopping, etc.

REFERENCES


ACKNOWLEDGEMENT

This work has been conducted by 1-year thesis activities in Computer Science and Engineering Discipline of Khulna University, Bangladesh, 2008-2009. Both the Lecturers of Asian University of Bangladesh and Khulna University have finalized the research paper.

BIOGRAPHY

G.M. Atiqur Rahaman received B. Sc. in Computer Science and Engineering with Distinction from Computer Science and Engineering Discipline in 2003. Currently he is Assistant Professor of Computer Science and Engineering Discipline of Khulna University, Bangladesh. He has published about 08 papers in national and international journals as well as in international conference proceedings. His areas of interest include Machine Intelligence, Data Mining, Computer Vision, Digital Image Processing, Color Informatics etc.

Md. Ahsan Arif has obtained B. Sc. in Computer Science from Computer Science and Engineering Discipline in 1999 and Master in Computer Application in 2001. Currently, he is a Senior Lecturer of Computer Science and Engineering Department of Asian University of Bangladesh (AUB). He has published about 03 papers in national and international journals. His areas of interest include Machine Intelligence, Data Mining, Computer Vision, Digital Image Processing, UNL, Compiler Construction, etc.

Eyrin Chowdhury appeared B.Sc. in Computer Science and Engineering (CSE) from Khulna University, Bangladesh in 2009. Her areas of interest include Computer programming, computer security, data mining, database system, operating system and Web software development.
Sharmi Debnath appeared B.Sc. in Computer Science and Engineering (CSE) from Khulna University, Bangladesh in 2009. Her areas of interest include Operating system, computer security, data mining, database system and software development.