International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

E-Mine: A Web Mining Technique for Relevant Data Regions

Vaishnavi J. Deshmukh[1], Anuja K. Pande[2], Sapna.Kaushik[3], Pallavi B. Rathod[4], Payal.Pawade[5]

Abstract

In recent years government agencies and industrial enterprises are using the web as the medium of publication. Hence, a large collection of documents, images, text files and other forms of data in structured, semi structured and unstructured forms are available on the web. It has become increasingly difficult to identify relevant pieces of information since the pages are often cluttered with irrelevant content like advertisements, copyright notices, etc surrounding the main content. Thus, we propose a technique that mines the relevant data regions from a web page. This technique is based on three important observations about data regions on the web.

Keyword

MDR, eMine algorithm, container Identifier, Largest Rectangle Identifier,filter

1. Introduction

Extracting the regularly structured data records from web pages is an important problem. So far, several attempts have been made to deal with the problem. The main disadvantage with the existing automatic approaches is their assumption that the relevant information of a data record is contained in a contiguous segment of HTML code, which is not always true. Thus, we propose a more effective method to mine the data region in a web page. The algorithm, eMine, finds the data regions formed by all types of tags using visual cues. Related work, mainly in the area of mining data records in a web page is MDR (Mining Data Records). MDR is a well-known approach which basically exploits the regularities in the HTML tag structure directly. MDR algorithm makes use of the HTML tag tree of theweb 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.

MDR 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 [1]. It has several other limitations which will be discussed in the latter half of this paper.

The algorithm is based on two observations:

(a)A group of data records are always presented in a contiguous region of the web page and are formatted using similar HTML tags. Such region is called a Data Region.

(b) The nested structure of the HTML tags in a web page usually forms a tag tree and a set of similar data records are formed by some child sub-trees of the same parent node. MDR system is a freeware and can be downloaded at:
http://www.cs.uic.edu/~liub/WebDataExtraction/MDR-download.html

2. The Proposed Technique

We propose a novel and an effective method, eMine, to mine the data region from a web page automatically. The basic criteria which eMine uses are the locations on the screen at which tags are rendered i.e. visual Information.

These help the system in three ways:

a) It enables the system to identify gaps that separate records, which helps to segment data records correctly, because the gaps within the data record(if any) is typically smaller than that in between data records.
b) The visual information also contains information about the hierarchical structure of the tags.
c) By observing a webpage, it can be analyzed that the relevant data region occupies the major central part of the Webpage.

The system model of the eMine technique is shown in fig 1. It consists of the following three components:

- Largest Rectangle Identifier.
- Container Identifier.
- Filter

The output of each component is the input of the next component.

The eMine technique is based on three observations:

- A group of data records, that contains descriptions of set of similar objects, is typically presented in a contiguous region of a page.

- The area covered by a rectangle that bounds the data region (refer to definition 1 below) is more than the area covered by the rectangles 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 [3].

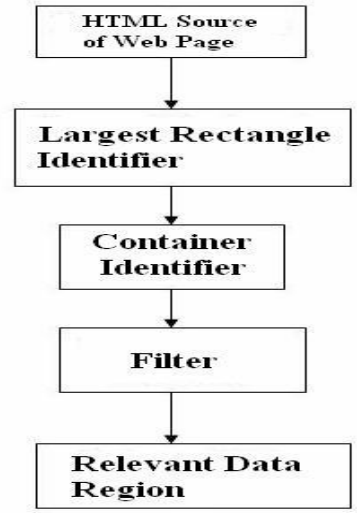

Figure 1: System model

Definition 1: A data region is defined as the most relevant portion of a webpage.
Definition 2: A data record is defined as a collection of data. It is a meaningful independent entity.
E.g. A product listed inside a data region on a product related web site is a data record.
Fig.2 illustrates an example which is a segment of a webpage (www.amazon.com) that shows a data region. The full description of each book is a data record.
Definition 3: For each tag, there exists an associated rectangular area on the screen. This rectangle is called the bounding rectangle for the particular tag.
The overall algorithm of the proposed technique is as follows [4]:
AlgorithmeMine
Input: The HTML source of the Web Page.

1. Determine the height & width of all the bounding Rectangles in the HTML document.
2. Calculate the areas of all the Bounding Rectangles.
3. Identify the Maximum Rectangle from all the bounding Rectangles.
4. Identify the container within the Maximum Rectangle obtained from step 3.
5. Identify the Data Region in the container obtained from step 4[4].

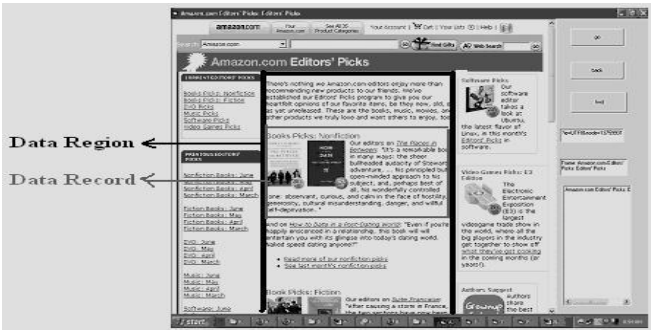

Figure 2:An Example of a page showing data region and data record (shown from eMine.exe)

3. AlgorithmeMine

Input: ***The HTML source of the Web Page.***

1. Determine the height & width of all the bounding Rectangles in the HTML document.
2. Calculate the areas of all the Bounding Rectangles.
3. Identify the Maximum Rectangle from all the bounding Rectangles.
4. Identify the container within the Maximum Rectangle obtained from step 3.
5. Identify the Data Region in the container obtained from step 4.

Working of an Algorithm*:*
The algorithm takes the HTML source of the web page as input. In step 2 we scan the HTML document for tags and identify the height and width of all the bounding rectangles. Thus, you have the area of each bounding rectangle. The step 3 finds the largest rectangle out of all the bounding rectangles. Step 4 identifies the container which holds most of the relevant data region (and some irrelevant regions also). Step 5 identifies the actual relevant data region by filtering the irrelevant regions. The following sections provide more details about the individual modules associated with the algorithm [5].

3.1Determining the Height and Width of all Bounding Rectangles

In the first step of the proposed technique, we determine the dimensions of all the bounding rectangles in the web page. Every <table> tag in a web page will be associated with a specific height

and width attribute. We extract them. If not specified, the MSHTML parsing and rendering engine of Microsoft Internet Explorer 6.0 can be used. This parsing and rendering engine of the web browser gives us the coordinates of a bounding rectangle. We scan the HTML file for tags. For each tag encountered, we determine the coordinates of the bounding rectangle of the corresponding tag and plot it.

The Fig. 3 shows a sample web page of the product related website, which contains a number of books; and their description which form the data records inside the data region [6].

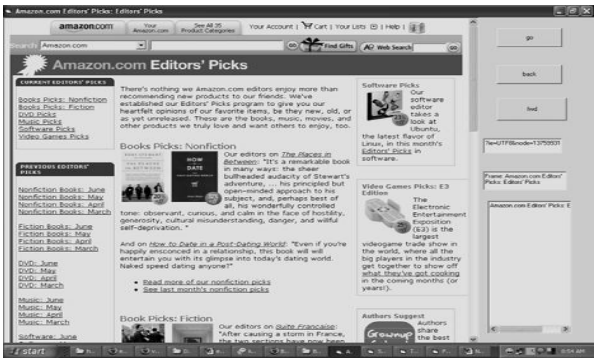

Figure 3: A Sample Web page of a product related website shown in eMine.exe

Fig 4 shows the bounding rectangles for the <td> tags of the web pages shown in Fig 3.

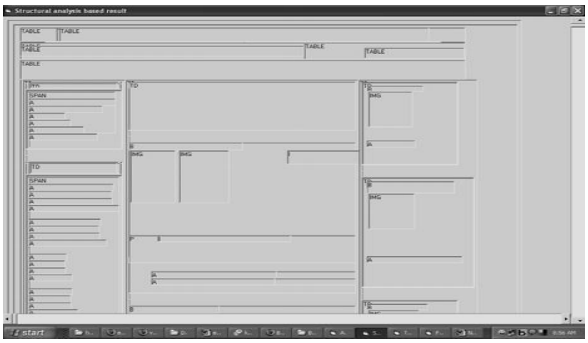

Figure 4: Bounding Rectangles for <TD> tag corresponding to the web page in Figure.3

3.2 Identification of the Largest Rectangle

Based on the height and width of bounding rectangles obtained in the previous step, we determine the area of the bounding rectangles of each of the children of the <body> tag. We then determine the largest rectangle amongst these bounding rectangles. The reason for doing this is a sensible assumption; that the largest bounding rectangle will always contain the most relevant data in that web page. The procedure followed to accomplish this task is as follows:

Procedure getMaxRect
Input: <body> of the HTML source
for each child of <body> tag
begin
Find the coordinates of the bounding rectangles for the child
If the area of the bounding rectangle > area of maximum Rectangle
then Maximum Rectangle = child
endif
end

3.3 Identification of the Container within the Largest Rectangle

Once we have obtained the largest rectangle, we form a set of the entire bounding rectangles. The rationale behind this is that the most important data of the web page must occupy a significant portion of the web page. Again, we determine the bounding rectangle having the largest area in the set. The reason for determining the largest rectangle within this set is that only the largest rectangle will contain data records. Thus a container (Refer to definition 4 below) is obtained which 'holds' the data region and also possibly, some irrelevant data [7].

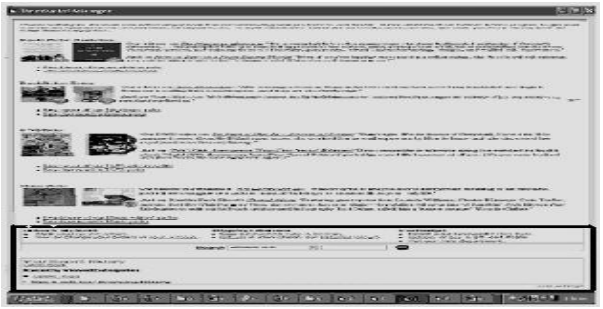

Figure 5: The container within the Largest Rectangle identified from sample web page in Figure 3.

A container is a superset of the data region which may or may not contain irrelevant data. For example, the irrelevant data contained in the container may include advertisements at the bottom of the page and followed by search bars or links to some other sites. The Fig.5 shows the container identified from the web page shown inFig.3. The procedure getContainer identifies the container in the web pages which contains the relevant data region along with some irrelevant data also. It is as follows:

Procedure getContainer

Input: The Largest Rectangle out of all Bounding Rectangles.
List_of_Children=depth first listing of all the children of the tag associated with Maximum Rectangle.
for each tag in List_of_Children
begin
if area of bounding rectangle of a tag > half the area of Maximum Rectangle
then container = tag
endif
end
The fig.6 shows the extracted regions from the container shown in fig.5. We note that there is some irrelevant data, at the bottom of the actual data region containing the data records.

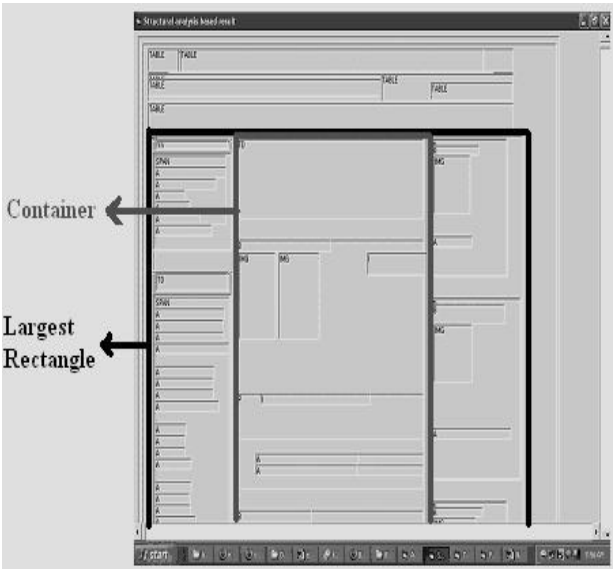

Figure 6: The extracted Regions from the container shown in Figure 5. The irrelevant portion to be filtered is highlighted.

3.4 Identification of Data Region Containing Data Records within the Container

To remove the irrelevant data from the container, we use a filter. The filter determines the average heights children within the container. Those children whose heights are less than the average height are identified as irrelevant and are discarded. The fig.7 shows a filter applied on the container in fig.6, in order to obtain the data region.
The procedure Filter filters the irrelevant data from the container, and gives the actual data region as the output [8]. It is as follows:
Procedure Filter
Input: The container obtained from the previous step.
totalHeight=0
for each child tag within container
totalHeight+=height of the bounding rectangle of child
averageHeight = totalHeight/no of children of container
for each child within container
if height of child's bounding rectangle <averageHeight
then Discard child from container
endif
end for
end for

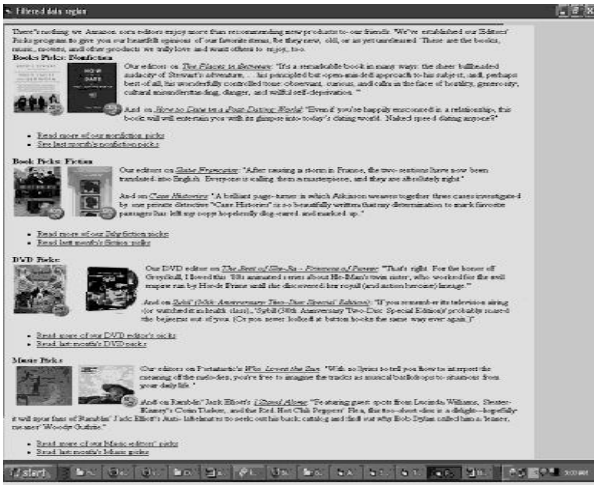

Figure 7: Data Region obtained after filtering the container in Figure 6.

Thus, the eMine technique, as described above, is able to mine the relevant data region containing data records from the given web page efficiently.

4. MDR VseMine

In this section we evaluate the proposed technique and also compare it with MDR. The evaluation consists of three aspects as discussed in the following

4.1 Data Region Extraction

We compare the first step of MDR with our system for identifying the data regions. MDR is 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> etc. In the proposed eMine system, the data region identification is independent of specific tags and forms. Unlike MDR, 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 of eMine, because the hierarchy of tags is constructed based on the visual cues on the web page. In case of MDR, the entire tag tree needs to be scanned in order to mine data regions, but eMine scans only the largest child of the <body> tag. Hence, this improves the time complexity compared to MDR.

4.2 Data Record Extraction

MDR identifies the data records based on keyword search (e.g. "$"). But eMine does not make use of any text or content mining. This proves to be very advantageous as it overcomes the additional overhead of performing keyword search on web page. MDR, not only identifies the relevant data region containing the search result records but also extracts records from all the other sections of the page, e.g. some advertisement records also, which are irrelevant.
In MDR, comparison of generalized nodes is based on string comparison using normalized edit distance method. However, this method is slow and inefficient as compared to eMine where the comparison is purely numeric.It scales well with all the web pages[9]

4.3 Overall Time Complexity

The existing algorithm MDR has complexity of the order O(nk) without considering string comparison, where n is the total number of nodes in the tag tree and k is the maximum number of tag nodes that generalized node can have (which is normally a small number <10). Our algorithmeMine has a complexity of the order of O (n), where n is the number of tag-comparisons made [10].

5. Conclusion and Future Work

In this paper, we have proposed a new approach to extract structured data from web pages. Although the problem has been studied by several researchers, existing techniques make many strong assumptions. eMine is a pure visual structure oriented method that can correctly identify the data regions. Most of the current algorithms fail to correctly determine the data region, when the data region consists of only one data record. Also, most of the approaches fail in the case where a series of data records is separated by an advertisement, followed again by a single data record. eMine works correctly for the above case. Further, the comparisons are made on numbers, unlike other methods where strings or trees are compared. Thus eMine overcomes the drawbacks of existing methods and performs significantly better than existing methods.

Extraction of the data fields from the data records contained in these mined data regions can be considered in the future work taking also into account the complexities such as the web pages featuring dynamic html, etc. 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, e.g., making a comparative study of products from various companies, smart shopping, etc.

References

[1] Bing Liu, RobertGrossman, YanhongZhai."Mining Web pages for Data Records", IEEE Computer Societ, pp. 49-55Nov/Dec 2004.
[2] Jiawei Han and MichelineKambler, "Data Mining: Concepts and Techniques", Diane Cerra San Francisco, CA 94111, pp 77-82, 2006.
[3] Arun .K. Pujari. "Data Mining Technique", International Journal of Advanced Research inComputer Science and Software Engineering, India, Vol 2, Issue 12, pp. 439-442Oct.2012.
[4] PieterAdriaans, DolfZantinge, "Data Mining".
[5] George M. Maracas, Modern Data Warehousing, Mining, and Visualization Core Concepts, 2003.
[6] J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo, Extracting semi-structured information from the web.
[7] A. Arasu, H. Garcia-Molina, Extracting structured data from web pages.
[8] C. Chang and S. Lui. IEPAD: Information extraction based on pattern discovery. In *Proc. of 2001 Intl. World Wide Web Conf.*, pages 681–688, 2001.
[9] S. Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th Intl. Conf. on Extending Database Technology*, 1998.
[10] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J.Widom, "The TSIMMIS Project: Integration of Heterogeneous Information Sources," InProceedings of Tenth Anniversary Meeting of the Information Processing Society of Japan, Tokyo, Japan, 7-18, 1994.

Ms. Vaishnavi J Deshmukh, 3 June'1989, received her bachelor's degree in Computer Science& Engineering from Amravati University and currently pursuing her master's. She is currently working as an Assistant Professor with the Department of Computer Engineering, Dr. Bhausaheb Nandurkar College of Engineering and Technology, Yavatmal, S.G.B.A.U-Amravati University,India. Her research interests mainly focused on E-Commerce, Mobile Ad hoc Networks, and Network Security.

Ms. Anuja K. Pande,27April'1989, received her bachelor's degree in Computer Science& Engineering from Amravati University and currently pursuing her master's. She is currently working as an Assistant Professor with the Department of Computer Engineering, Dr.BhausahebNandurkar College of Engineering and Technology, Yavatmal, S.G.B.A.U-Amravati University, India.Her research interests mainly focused on E-Commerce, Mobile Ad hoc

SapnaS.Kaushik, 11 June'1973, received her bachelor's and master's degree in Computer Science& Engineering from Amravati University. She is currently working as a Head of the Department of Computer Engineering, Dr .BhausahebNandurkar College of Engineering & Technology-Yavatmal, S.G.B.A.U-Amravati University, India.
Her research interests mainly focused on, Mobile Ad Hoc Networks, and Network Security, E-Commerce.

Ms. Pallavi B. Rathod, 27 April 1989, received her bachelor's degree in Computer Science& Engineering from Amravati University and currently pursuing her master's from Sipna College of Engineering & Technology, Amravati. Her research interests mainly focused on Network Security.

Ms. Payal. Pawade,29August'1989, received her bachelor's degree in Computer Science& Engineering from Amravati University and currently pursuing her master's from Sipna College of Engineering & Technology, Amravati. Her research interests mainly focused on Network Security.