Visual Clue Based Extraction of Web Data from Flat and Nested Data Records

Siddu P Algur
Dept of Info Sc & Engg
SDM College of Engg & Tech, Dharwad
Karnataka ,India
siddu_p_algur @hotmail.com

P S Hiremath
Dept of Computer Science
Gulbarga University, Gulbarga
Karnataka ,India
hiremathps@yahoo.co.in

Abstract

This paper studies the problem of identification and extraction of structured data items from the nested and flat records of given web pages. Each of such pages may contain several groups of structured records. Most of the existing methods still have certain limitations. In this paper, we propose a more novel and effective technique for the extraction of data items. Given a page, the proposed technique first identifies the data region based on the visual clue information. It then extracts each record from the data region and identifies it whether it is a flat or nested records based on visual information – the area covered and the number of data items present in each record. The next step is data items extraction from these records and transferring them into the database. Once the data items are present in the database knowledge discovery can be carried out. This technique extracts data items fro the both nested and flat records. Our experimental results show that the proposed technique is effective and better than existing techniques.

Categories and Subject Descriptors

Data Warehousing and Mining, Web Services

Keywords

Web mining, Web data regions, Web data records

1. Introduction

More and more companies manage their business and publish their products and services on the Web. Collecting and organizing these dynamic information can produce the data for many value-added applications. In order to collate and compare the prices and features for products from the various Web sites, we need tools to extract attribute descriptions of each product (called data object) within a specific region (called data region) in pages.

International Conference on Management of Data
COMAD 2006, Delhi, India, December 14-16, 2006
©Computer Society of India, 2006

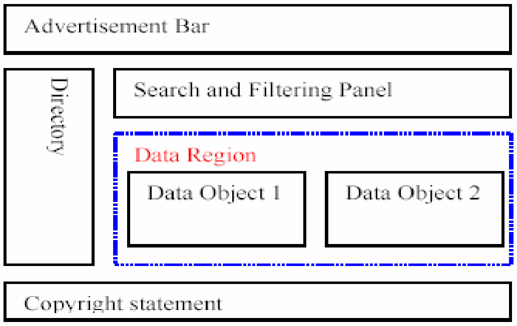


Fig 1: A schematic view of a webpage

As shown in Fig. 1, There are many irrelevant components intertwine with the descriptions of data objects in Web pages. These items include advertisement bar, product category, search panel, navigator bar, and copyright statement. In many Web pages, there are normally more than one data object intertwined together in a data region. Furthermore, the raw source of the Web page for depicting the objects might be non-contiguous. So it is difficult to discover the attributes for each object.

In real applications, what the users want from complex Web pages is the description of individual data object derived from the partitioning of data region. The approaches [2],[3],[4],[5],[8] have been proposed in literature to address the problem of web data extraction, which is also called wrapper generation.

This paper proposes a novel and more effective method to extract data items in a web page automatically. The model is called VCED (**V**isual **C**lue based **E**xtraction of web **D**ata from flat and nested data records) which is a list based approach. It finds the data items formed by all types of tags.
Given a web page, it works in two main steps.
i) Identification and Extraction of data region based on visual clue [location of data region/ data records/ data items on the screen at which the tags are rendered] information of web pages.
ii)Identification of data records and extraction of data items from it.
Experimental evaluation shows that the technique is highly effective.

2. The Proposed System Model

The proposed model is called **VCED** which is an extension of VSAP[10] technique, which extracts the data region from a given web page.
The system model of the **VCED** technique is shown in Fig 2. It consists of the following components.

1. Extraction of data records
2. Identification of data records
3. Extraction of data fields

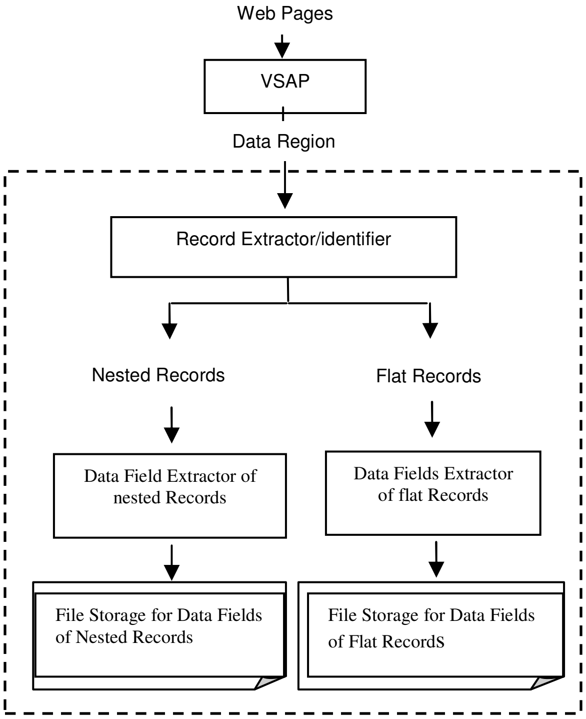

Web Pages
VSAP
Data Region
Record Extractor/identifier
Nested Records
Flat Records
Data Field Extractor of nested Records
Data Fields Extractor of flat Records
File Storage for Data Fields of Nested Records
File Storage for Data Fields of Flat Records

Fig. 2 System Model

The output of each component is the input of for the next component
Fig 3 illustrates an example which is a segment of web page that shows flat and nested data records.

Definition 1: A *flat data* records is defined as a collection of data items that together represents a single meaningful entity.
i.e., the product having single size, look, price etc.,
Definition 2: A *nested data* is defined as one that provides multiple description of the same entity.
i.e., the same type of products but different sizes, looks, prices etc.,

2.1 Extraction of data records

Extraction of data records is based on visual clues. In the first step of the proposed technique, we determine the height of all the data records. This approach uses the MSHTML parsing and rendering engine that gives the height of each data record. The height of the data record is obtained from the offsetHeight property of the HTMLObjectElement. Next the average height of the records is calculated. The average height of all the records provides the approximate height of each record The height of each data record is compared with the average height, if the height of the child is greater than or equal to the average height then the data record is extracted.

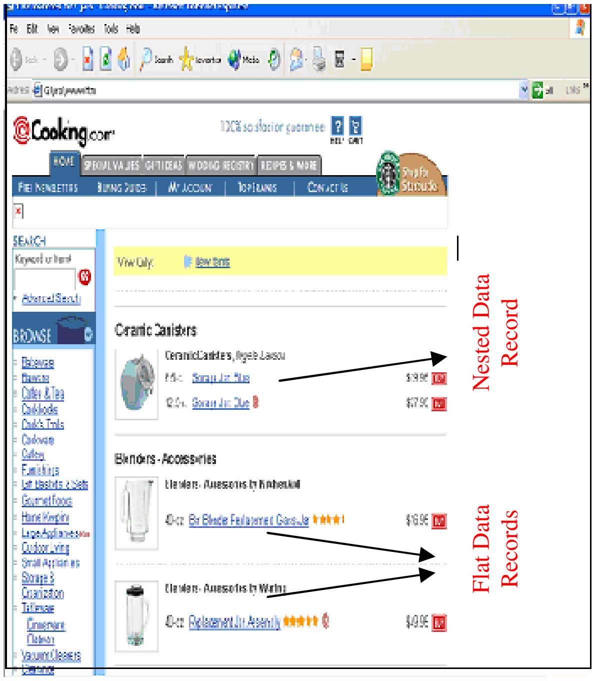
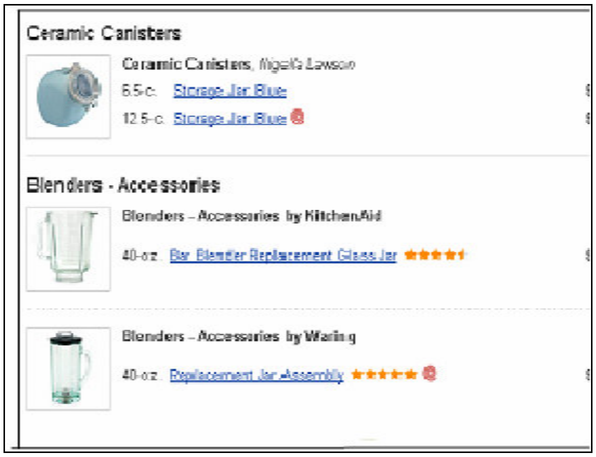


Fig. 3 An example of Flat and Nested data records

The procedure Extract Data Record, extracts the flat and nested records from given data region. It is as follows.



Ceramic Canisters

Ceramic Canisters, Nigella Lawson

6.5-c Storage Jar Blue

12.5-c Storage Jar Blue

Blenders - Accessories

Blenders – Accessories by KitchenAid

40-oz Bar Blender Replacement Glass Jar ★★★★

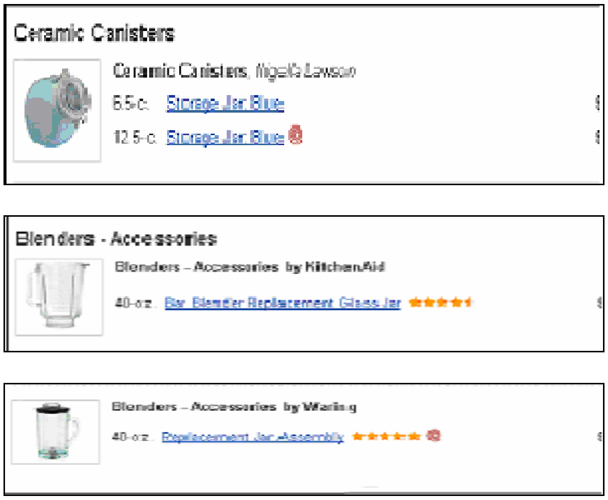
Blenders – Accessories by Waring

40-oz Replacement Jar Assembly ★★★★★

Fig. 4 Filtered data region

```
Procedure ExtractDataRecord(dataRegion)
{
   THeight=0
     For each child of dataRegion
       BEGIN
          THeight += height of the bounding
                      rectangle  of child
       END
        AHeight = THeight/no of children of
                  dataRegion
     For each child of dataRegion
       BEGIN
          If height of child's bounding rectangle >
          AHeight
         BEGIN
           dataRecord=child
         END
       END
}
```

Fig. 5 shows extracted data records from the data region of Fig.4.



Ceramic Canisters

Ceramic Canisters, Nigella Lawson

6.5-c. Storage Jar Blue

12.5-c. Storage Jar Blue

Blenders - Accessories

Blenders – Accessories by KitchenAid

40-oz. Bar Blender Replacement Glass Jar ★★★★★

Blenders – Accessories by Waring

40-oz. Replacement Jar Assembly ★★★★★

Fig .5 Extracted data records

2.2 Identification of data records

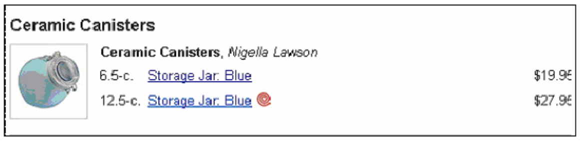
Identification of data records as flat or nested is essential in order to simplify the task of extracting the data items, which is very much needed for various applications as mentioned.

This technique determines the data fields for each data record within the data region based on the tags. Various tags such as <TD>, <TR>, <A>, <IMG>, represent the data fields. By counting these tags as they are encountered the number of fields is obtained. The flat record gives description of a single entity whereas the nested data record gives multiple description of a single entity, so the data fields in flat records are less as compared to that of nested records. Experimental observations have proved that the number of fields in the nested data records is approximately minimum of 40% more than that of the flat records. The number of fields in the first record is compared with the number of fields in the next record. If the number of fields is more than 40% then it is a nested record else it is a flat record. Suppose a condition is encountered where the number of fields is equal then determining whether the record is flat or nested becomes ambiguous. In order to overcome this problem the record is compared with the third record and so on till the condition is sufficed.

The procedure identify nested data record, identifies whether the record is flat or nested based on the number of data items present in the data record.

```
Procedure IdentifyNestedData(dataRecord[I],
dataRecord[I+1])
{   noofField[I]=0
      For I 1 to no of records
         BEGIN
           noofFields [I]=noofFields[I]+noofFields in the
           record[I]
         END
      DO
           For I 1 to no of records
        BEGIN
           For dataRecord [I], dataRecord[I+1]
           IF the no of fields in the [I+1] th record>=40%
           of the no of fields in the [I] th record
           The [I+1]th record is a nested data record
          ELSE
           The [I] th record is a nested data record
         END
      WHILE (EOF)
 }
```

In Fig. 6a and Fig. 6b shows the identified nested and flat data records. In Fig.6a the number of data fields is 12 and in Fig.6b the number of data fields is 7. The number of data fields in Fig 6a is 58.3% more than the number of data fields in Fig 6b. From this observation it is clear that 6a is nested data record and 6b is flat data record.



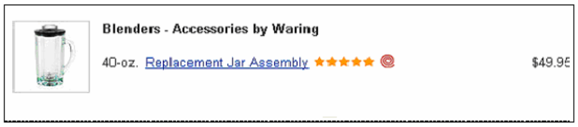
Ceramic Canisters

Ceramic Canisters, *Nigella Lawson*

6.5-c. Storage Jar: Blue $19.9

12.5-c. Storage Jar: Blue $27.9

Fig 6a: Identified nested data record
No. of data fields = 12



Blenders - Accessories by Waring

40-oz. Replacement Jar Assembly ★★★★★ $49.9

Fig 6b: Identified flat data record
No. of data fields = 7

2.3 Extraction of data fields from the extracted records.

Once the record is being extracted and identified the next step is to extract the data fields from the data records. The data fields are extracted based on the following algorithms.

```
Procedure ExtractNesteddatafields()
{
        extract nested records from Flatdata file.
        For I From the start of the file to the END of
        file
    BEGIN
        Extract the data fields row by row
      END
        Store the data fields in the file.
}
```

The above algorithm explains how data fields are extracted from nested records. First the file in which the nested data records are stored is navigated. The file is navigated using the absolute path of the file. Then the file is read line by line till the end of file. The data fields are extracted row by row. Each data field has a bounding rectangle associated with it. The data fields are extracted using these bounding rectangles. When a bounding rectangle is recognized the respective data field is extracted and stored in a file.

```
ProcedureExtractFlatdatafields()
{
        extract nested records from Flatdata file.
        For I From the start of the file to the end
    BEGIN
        Extract the data fields row by row
      END
        Store the data fields in the file.
}
```

The above algorithm explains the extraction of data fields from the extracted and identified flat records. The procedure for extracting the data fields from flat records is same as mentioned above for the nested records.

3. Conclusion:

In this paper we proposed a more effective and better technique to perform the automatic data extraction from the flat nested data records from the web pages. Given a web page our method first identifies and extracts the data records based on the visual clue information. It than counts the number of the data items in the each records and then identifies it as either flat or nested. The extracted data fields are then stored in file. Although the problem has been studied by several researchers, existing techniques either inaccurate or make many strong assumptions. The VCED is a pure visual clue based extraction of flat and nested data records.

4. References:

[1] Baeza Yates, R. Algorithms for string matching: A survey. ACM SIGIR Forum, 23(3-4):34—58, 1989.

[2] J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo . Extracting semi-structured information from the web.In Proc.of the Workshop on the Management of Semi-structured Data, 1997.

[3] Kushmerick, N. Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence, 118:15-68, 2000. Clustering-based Approach to Integrating Source Query]

[4] Chang, C-H., Lui, S-L. IEPAD: Information Extraction Based on Pattern Discovery. WWW-01, 2001.]

[5] Crescenzi, V., Mecca, G. and Merialdo, P. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. VLDB-01, 2001.]

[6] Liu, B., Grossman, R. and Zhai, Y. Mining Data Records in Web Pages. KDD-03, 2003.

[7] J. Wang, F. H Lochovsky. Data Extraction and Label Assignment for Web Databases.WWW conference, 2003.

[8] H. Zhao, W. Meng, Z. Wu, Raghavan, Clement Yu. Fully Automatic Wrapper Generation For Search Engines, International WWW conference 2005, May 10-14,2005, Japan. ACM 1-59593-046-9/05/005

[9] Zhai, Y., Liu, B. Web Data Extraction Based on Partial Tree Alignment , WWW-05, 2005, May 10-14, 2005, Chiba, Japan. ACM 1-59593-046-9/05/00

[10] S.S Benchalli, P.S Hiremath, Siddu Algur, Renuka Udapudi "Mining Data Regions from Web Pages" , COMAD2005b,2005, DEC.

[11] Bing Liu and Yanhong Zhai. "NET - A System for Extracting Web Data from Flat and Nested Data Records." Proceedings of 6th International Conference on Web Information Systems Engineering (WISE-05), 2005.

[12] Lerman, K., Getoor L., Minton, S. and Knoblock, C. Using the Structure of Web Sites for Automatic Segmentation of Tables. SIGMOD'04, 2004.