Copyright © 2016 American Scientific Publishers
All rights reserved
Printed in the United States of America

Advanced Science Letters
Vol. xx, 400–407, 2016

Approach for Unwrapping the Unstructured to Structured Data the Case of Classified Ads in HTML Format

Lintang Yuniar Banowosari,[1], Detty Purnamasari[2]

[1]Information Management Dept., Gunadarma University, Jakarta, Indonesia, Email: lintang@staff.gunadarma.ac.id
[2]Inforrmation System Dept., Gunadarma University, Jakarta, Indonesia, Email: detty@staff.gunadarma.ac.id

Data sources with various forms and formats available on the Internet. Data can be in the form of semi-structured and unstructured data. Research's objective is developing approach for unwrapping the unstructured data available on the internet into structured data / database. Unstructured data used in this study is in the case of classified ads on the Indonesia website, and those unstructured data is in HTML format. The Illustration made to test the approach. The results of the test show the value of f-measure 99.13%.

Keywords: Classified Ads, database, internet, unwrapping, unstructured data.

1. INTRODUCTION

Data can be obtained from various sources, one of which is the Internet that provide data with the various forms and formats. Data are presented on the website is designed to enable users and readers understand the data, but the data is may not necessarily well understood by the computer easily. One particular form of unstructured data is data on classified ads that are widely used in conducting the sale on the website in Indonesia. Website developers much more pay attention to how can display the data / information to the user / reader that easily understandable, regardless of how to make the data also understood by the computer.

The research's objective in this article is to develop an approach to extract unstructured data into structured data using classified ads car sales as an example of unstructured data in HTML format. Many researchers can use terminology extraction, converter or unwrapped. In the example the classifieds is not known which data as a property, and instance.

Property is the name of the column of the table and the instance / record / data contents of the cell is the table row and those data are related**.**

Unstructured data is data that does not have a formal structure and it does not mean the data is not displayed neatly but may be a way how establishing data / coding of the data is done in a way that is difficult to be implemented by computer, such as plain text. [5]. Unstructured data extraction is done to facilitate the use of these data for further processing, so that the results of this extraction can be combined with the other data.

2. RELATED WORKS

Research conducting unstructured data extraction was undertaken by [3] with the using of unsupervised learning algorithm to look at the structure of the list. Extraction of data in the form of lists and notice the template used to identify the list in each web page. Then calculates the features extracted from each data, which further identifying columns and rows.

1936-6612/2016/4/400/008 doi:10.1166/asl.2016.1261

Other studies have also done by generating an application that called Lixto [1] based on conversion from XML to HTML pages using a logical language called Elog. Lixto provide two (2) options to perform data extraction mechanism that is tree extraction to identify each element of the HTML tags and strings extraction to include elements of the cell contents into the HTML tag that has been formed by trees.

Wei et al [6] developed a technique to extract tables from the available text. These texts of such as a query performed on table. Activities carried out such as changing text constitute unstructured data into structured data, despite the fact that these structured data has been provided in the form of tables and just straight taken away. Currently there are many web sites are created using HTML, and the use HTML tags function to set the display on the browser. Data on the HTML hierarchy has two types, which are Syntactic Hierarchy and Hierarchy Intended.

Syntactic hierarchy is a hierarchy that is made with the attention to grammar in HTML (HTML tags), whereas Intended hierarchy is a hierarchy in the form of content / data presented in an HTML table. [4].

One example of syntactic hierarchy is the Document Object Model (DOM tree), that is the basic and the stand-alone language used to represent and make the connection between objects of different documents into HTML web pages with the form of tree structure as a depiction. [2]

Research this unstructured data; take the example of the data on the classified ads website made with HTML tag with respect to syntactic hierarchy.

3. METHOD

The form of unstructured data used in this study was classified advertisements. Classified ads contained on the website form of collection of data regarding a product / item being sold and the data are arranged to look like in one sentence and is on a single column.

Web site built using many <table> tags to adjust these website displays. So, in the extraction of unstructured data in this study is to look at <table> tag existing HTML tags.

Unstructured data extraction steps are shown in Figure 1. Extraction step of unstructured data into the database is as follows:

1. Pre Processing
2. Retrieving data from a web site that contains a table that will be extracted. Taking process (grab) HTML tags using a tool, and with these grabs results of HTML tag, extraction process carried out.
3. Extracting the HTML tag to get the candidate table containing classified ads by defining <table> ... </table> tag and perform a search to determine the number of rows with the <tr> ... </ tr> tag and the determination of the number of columns by <td> ... </td> tag.
4. Conducting the selection on the candidate table containing classified ads by matching words that exist in the database Keywords to find a table containing classified advertisements.
5. Matching data contained, after the table containing classified ads found, the next step is to match the data contained in the tables that contain classified ads with the tag word and data word.
6. Extracting results are stored in the database

Preprocessing step is carried out in sequence illustrate as follows:

a. Define the HTML tags, that is <table> ... </ table> tags
b. Search for the word most often used as a keyword in the classified ads, and stored on a Keywords database.
c. Looking for a data word, that is any data related to the products / goods sold in the classifieds.
d. Looking tag word that often appears on the classifieds to distinguish the words found in these classified ads whether a part of the information about the products / items advertised

The logic or algorithm to find table that contains advertising as follow:

1. If the first tag is found <html> tag, and in inside there is <table>…</table> tag, table attraction can be executed.
2. If there is <table> v^{th} tag, content of the tag is table candidate table that has advertising (set v^{th} start with value =1).
3. Read tag <tr> in <table>…<table> tag the value will be saved as number of candidate table.
4. Read <td>…<td> tag to consider number of column/cell to save other data in <td>…<td> tag
5. Repeat step 4 until find </tr> tag, continue to read <td>…<td> tag.
6. Repeat step 3 if not yet found tag <table>, continue to read <tr>…</tr> tag.
7.Repeat step 3 until found <html>…</html> tag and continue to read <table>…</table> tag.
8. After process extraction HTML tag finished, the number of <table> tag can be calculated as candidate table that consist advertising convert to structured data.

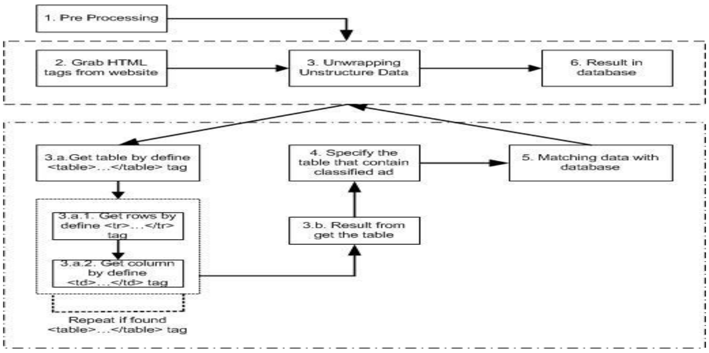

1. Pre Processing
2. Grab HTML tags from website
3. Unwrapping Unstructure Data
6. Result in database
3.a.Get table by define <table>…</table> tag
4. Specify the table that contain classified ad
5. Matching data with database
3.a.1. Get rows by define <tr>…</tr> tag
3.b. Result from get the table
3.a.2. Get column by define <td>…</td> tag
Repeat if found <table>…</table> tag

Figure 1. Steps of Unstructured Data into the Database

Algorithm 1. Extract Table as Candidates Classified Ad

```
Read HTML
TableQty = 0; RowQty = 0; ColQty = 0;
If read tag <table> then TableQty = TableQty + 1
Loop If;
For x = 1 to TableQty do
    Begin
      If read tag <tr> then RowQty = RowQty + 1
      Loop If ;
      For y = 1 to RowQty do
        Begin
         If  read tag <td>…</td> then ColQty = ColQty + 1
          Loop if;
        End ;
      End;
p = 1; q = 1; r = 1;
    For p =1 to TableQty
      For q = 1 to RowQty
        For r = 1 to ColQty
          TdVal (p,q,r) = value in tag {<td>} r
          Save TdVal (p,q,r) as element cell
        Next r;
      Next q;
    Next p;
```

Note:

1. TableQty: number of <table> tags used in HTML web site
2. RowQty: the number of existing rows in a table.
3. ColQty: number of columns / existing cell of each row of a table.
4. TdVal: the value that is in the <td> ... </ td> tag denote the data / elements of the cell.
5. p, q, r: auxiliary variable for looping.

If the classifieds candidates table has been extracted, the next step is to determine which table contains classified ads by means of matching words with the Keywords database. Rules for detection table containing classified ads are: Perform Keyword database matching on each classifieds candidate table which are result of <table> ... </ table> tag extraction.

If the number of words in the Keywords database that can be found in the classifieds candidate table amount equal to or greater than a threshold value (the number of words found => Threshold), it is said that these table is a table that contains classified ads, where the data is unstructured .

Formally, the determination of the number of words to define tables that contain classified ads shown below:

Classified ads= {Σ (Keyword → Table Candidate) => Thresholds}

Algorithm 2 is used to define a table as a classified ad is as follows:

Algorithm 2 is used to define a table as a classified ad is as follows:

```
SimWordQty  = 0; TH
For p = 1 to TableQty  do
    For q = 1 to RowQty do
      For r = 1 to ColQty do
        Read TdVAl (p,q,r)
        While s = 1 do
          Read recKeyWord (s)
          If RecKeyWord (s) = TdVal (p,q,r) then
            Begin
                SimWordQty = SimWordQty +1
                If  SimWordQty  >=TH then
                "{Table} p is classified ad"
                Else s = s + 1;
            End;
        Else s = s + 1 ;
        Until s = 8 ;
      Next r ;
    Next q ;
    " {Table} p isn't classified ad"
Next p;
```

Note:

1. TableQty: number of <table> tags used in HTML websites.
2. RowQty : the number of existing rows in a table
3. CokQty: existing number of columns of each row of a table
4. TdVal: the value that is in the <td> ... </ td> tag is the data / elements of the cell.
5. SimWordQty: the number of words in the database of keywords found in the tables extracted data.
6. RecKeyWord: keyword read from the keyword database
7. p, q, r, s: auxiliary variable for looping.

By using the algorithm 2, a table containing classified ad has been discovered, the next step is to change the unstructured data into classifieds structured data / database. How that is done by using the tag word and data word. Tag word made in 2 (two) categories, that is: tag word differentiator word as part of the data on products/goods to the classified ads, and tag word differentiator word for ignoring strings/words after the tag word. The steps are as follows:

1. Match the tag word on the database containing classified advertisements.
2. If the tag word is not found and there is still a tag word being read, then read the next tag word, repeat step 1. If the tag word is not found and all of the tag word has been read, then do the second way/method by performing the matching of classified ads database with the data word.
3. If the tag word found in the classifieds database, grab the string that exist in the classifieds database after tag word as much as the value of the number maximum words that matched these tag word.
4. String as much as the number maximum words values

taken after the tag word matched with the data word stored in the database corresponding with the tag word.

5. Save string matching results according to the data word which is found in structured data storage, and use flags to mark off the word in the next classifieds database that will be matched.
6. If still there tag word that has not been matched, then repeat step 1 for the next tag word.
7. If the all of the tag word matched is finished, then do the second way/method that is doing matching of classifieds database with the data word.

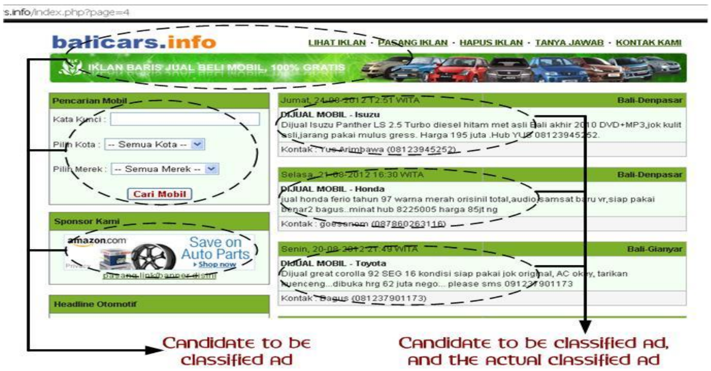
After performing the matching of tag word on classified ad database, the next step is to perform matching of data. In the matching process classified ads database with the data word per category, are also made to match the tag word which is said word differentiator to ignoring a number of strings / words after these tag word. Steps are as follows:

1. Data word per category matched on a database containing classified advertisements.
2. If not found and still no data word to be matched, and then move on to the next data word, repeat step 1.
3. If the data word that matched was found, then grab the string before the word that fits these data word. Number of string is taken as much as the threshold value.
4. Match these tag word with the number of string taken from step 3.
5. If the tag word is not found in number of string which taken, then save the word/string that are matched to database of structured data.
6. If you still have the data word per category which will be matched, then read the next data word, and repeat step 1.
7. If the tag word was found in number of string taken, then ignore the word of matching results with the data word.
8. If you still have data word per category which will be matched, then read the next data word, and repeat step 1.

4. RESULT AND DISCUSSION

Based on approach of the extraction of unstructured data into structured data in this study, is made illustrations for the extraction of unstructured data, which is about the classified ad selling a car that is on the Indonesia website. Figure 2 is an example of classified ads car sales available on the Indonesia website.

Pre-processing stage in order to create a keywords database for detection the table on extract HTML tag is to do a survey on 30 web sites which provide classified ads and see what the word which widely used.



s.info/index.php?page=4

balicars.info

LIHAT IKLAN · PASANG IKLAN · HAPUS IKLAN · TANYA JAWAB · KONTAK KAMI

IKLAN BARIS JUAL BELI MOBIL, 100% GRATIS

Pencarian Mobil

Kata Kunci :

Pilih Kota : -- Semua Kota --

Pilih Merek : -- Semua Merek --

Cari Mobil

Sponsor Kami

amazon.com Save on Auto Parts ▸ Shop now

pasang link/banner disini

Headline Otomotif

Jumat, 24-08-2012 12:51 WITA Bali-Denpasar

DIJUAL MOBIL - Isuzu

Dijual Isuzu Panther LS 2.5 Turbo diesel hitam met asli Bali akhir 2010 DVD+MP3,jok kulit asli,jarang pakai mulus gress. Harga 195 juta .Hub YUS 08123945252.

Kontak : Yus Arimbawa (08123945252)

Selasa, 21-08-2012 16:30 WITA Bali-Denpasar

DIJUAL MOBIL - Honda

jual honda ferio tahun 97 warna merah orisinil total,audio,samsat baru vr,siap pakai benar2 bagus..minat hub 8225005 harga 85jt ng

Kontak : goessanom (087860263116)

Senin, 20-08-2012 21:49 WITA Bali-Gianyar

DIJUAL MOBIL - Toyota

Dijual great corolla 92 SEG 16 kondisi siap pakai jok original, AC okey, tarikan kuenceng...dibuka hrg 62 juta nego... please sms 081237901173

Kontak : Bagus (081237901173)

Candidate to be classified ad
Candidate to be classified ad, and the actual classified ad

Figure 2. Example of Classified Ads of Cars Selling
Source : http://iklan.balicars.ifo/ [7]

This survey is also obtained Thresholds for determining the value of a classified ad is equal to or greater than 4 (TH => 4). These results presented in Table 1.

Table 1.Keywords Database
to Detect Table Classified Ads

No.	Words	No.	Words
1	*Jual/Dijual* (Sale)	5	*Hub/Hubungi* (Contact)
2	*Mobil* (Car)	6	*Nego* (Negotiable)
3	*Harga* (Price)	7	*Tahun* (Year)
4	*Warna* (Color)	8	*Kondisi* (Condition)

Then tag word is divided into 2 (two), that is word differentiator tag word as data about the car and the tag word as word differentiator to be ignored, and each tag word be given Maximum Value Word. Tag word is presented in Table 2 and Table 3.

Table 2.Tag Word Differentiator Words as a
Data about Cars

No.	Words	Number of the maximum words
1	*Tahun/Thn/Th* (Year)	2
2	*Warna/Wrn* (Color)	3
3	*Tipe/Type* (Type)	3
4	*Harga/Hrg/Rp* (Price)	3
5	*Merk/Merek* (Brand)	3

If the tag word classified ad found in Table 2 then the string after these tag word is the data which will be stored as structured data / database in accordance with its name of the tag word.

Table 3.Tag Word differentiator Word to be ignored

No.	Words	Number of the maximum words
1	*Hubungi/Hub* (Contact)	3
2	*Jalan/Jln/Jl* (Street)	3
3	*Kontak/Contact* (Contact)	3

Determination of the number of string after tag word which converted into structured data using number of the maximum word of each of these tag word, and then perform string matching with the data word. Tag word in table 3 used when making changes unstructured data into structured data with the data word, as shown in Figure 5. Websites with HTML tags that contain classified ads extracted, and executed the algorithm 1 and algorithm 2 to take unstructured data in these form of classified ads.

Once classified ads data obtained, then to convert it into structured data in the case of classified ads car sales using the following rules:

1. If found the word: "*Tahun/Thn/Th*", means Year then the next string is year Data of the car which sold.
2. If found the word: "*Warna/Wrn*", means Color the next string is data about the color of cars sold.
3. If found the word: "*Harga/Hrg/Rp*", means Price then the string after taking into account the type is numerical is data on the price of the car sold.
4. If found the word: "*Tipe / Type*", means Type the next string is Data about the type of car being sold.
5. If found the word: "*Merk/Merek*", means Brand the next string is data about brand cars sold.

Then rule on differentiator tag word says to ignore the string after the word tag is when found the word: "*Kontak/Contact*" means contact, "*Hubungi/Hub*" means contact, and "*Jalan/Jln/Jl*" means street, then the next string not included in the car data. Number of string after tag word which is been ignored as much as the number of minimum words. Data word prepared on preprocessing is divided into four (4) categories, which are: Category of car brands, Categories of car types, Category of car colors, Category of car year. Data word to category brand, type, color, and year of the car following the steps as shown in Figure 5 to make changes to classified ads unstructured data into structured data car sales.

An illustration that conducted by comparing the classified ad with the tag word contained in Table 1. On the particular web page was found that there is the word: "*warna*" means color in the classified ads, it is taken as the number of words after the word "*warna*" of a number minimum the word is 3 words: "*merah orisinil total*" means totally original red . These three words are matched with a database '*warna*", and *warna* is obtained "*merah*" means red, so "*merah*" is stored in a database (structured data) as data contents of the color property ("*warna*" database). After using the tag word, then executed a second way with the use the data word, which is matching word on the database four categories data of car, so that unstructured data is converted into structured data.

Illustration by comparing data word to the type of car category, which was found, is Honda. Then the word before Honda that is word sale is taken to be matched with the tag word to word differentiator is ignored. The result is not found, so the Honda is stored into the database as the storage of the data content property car type. The process of comparing Data word continued when there are other categories which have not been compared.

The approach in this study testing on 100 classified ads in Indonesian language contained in the 14 pages of the website. The algorithm is implemented in the PHP programming language. Testing was conducted to determine the ability of algorithms 1 and 2 in determining classified ads candidates and determine classified ads by these candidates, and also to determine the ability in perform the extraction of unstructured data in the case of classified ads car sales into structured data / database with the determine the contents of the property and property / instance of these classified ads. Test results are measured by calculating precision, recall, and F-measure. F-measure value obtained was 99.13%.

5. CONCLUSIONS

Extraction which carried out for unstructured data on classified ads into structured data performed in the case of classified ads which are found on the website of Indonesia. Previous research conducted by Wei e.al. [6] Which was begins with the text, then the text can be extracted from the properties of an existing table to create a new table extraction results. Test results [6] to determine these property values showed 95.8% f-measure, and the test results of this approach in this study showed better accuracy than other researchers. Unstructured data extraction capabilities into structured data / database on cases classified ads in this study is influenced by the completeness of Data Word and the Tag Word prepared in preprocessing.

REFERENCES

[1] Baumgartner, R., Flesca, S., Gottlob, G. 2001. Visual Web Information Extraction with Lixto. *Proceedings of the 27th International Conference on Very Large Data Bases*. Page 119 – 128

[2] Gultom, R., Sari, R. F., Budiardjo, B. 2011. Proposing the new Algorithm and Technique Development for Integrating Web Table Extraction and Building a Mashup. *Journal of Computer Science 7 (2)*. Page 129-142

[3] Lerman, K., Knoblock, C., Steven, M. 2001. Automatic Data Extraction from Lists and Tables in Web Sources. *Proceedings of Automatic Text Extraction and Mining Workshop*

[4] Lim, S. J., Ng, Y.K. 1999. An Automated Approach for Retrieving Hierarchical Data from HTML Tables. *Proceedings of The Eighth International Conference on Information and Knowledge Management*. Page 466 - 474

[5] Shaker, M., Ibrahim, H., Mustapha, A., Abdullah, L. N. 2008. A Framework for Extracting Information from Semi-Structured Web Data Sources. *Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology*. Page 27-31

[6] Wei, X., Croft, B., McCallum, A. 2006. Table Extraction for Answer Retrieval. *Journal Information Retrieval. Volume 9 Issue 5*. Page 589 – 611

[7] http://iklan.balicars.ifo/101 (2004), 3495-3497

doi:10.1166/asl.2016.1261