HTML Format Tables Extraction with Differentiating Cell Content as Property Name

*Detty Purnamasari[1], Lintang Yuniar Banowosari[2], I Wayan Simri Wicaksana[3], Suryadi Harmanto[4]

[1]Information System, Gunadarma University, Jl. Margonda Raya No. 100 Pondok Cina Depok, Indonesia
[2]Information Management, Gunadarma University, Jl. Margonda Raya No. 100 Pondok Cina Depok, Indonesia
[3,4]Information Technology, Gunadarma University, Jl. Margonda Raya No. 100 Pondok Cina Depok, Indonesia

Website presents data in various forms and formats, one of them in the form of a table. Tables on the Internet can be taken such way by copy and paste, but this way is not easy if done on many tables then from extracted result they have been merged with the other tables. This article discussed the research on extraction of HTML tables which stored into a database form. The approach used was algorithm to perform the search process the number of rows and number of columns from the table, and algorithms to perform matching the contents of the table cell extraction results with a Property Name database, so it is unknown whether the extracted table has property in the row/column/table without property. Table and Property Name database displays the data in the Indonesian Language. At pre processing stage Property Name database which is also prepared the techniques to enrich the instance of the Property Name database. The tables in the extract is a table HTML format with a simple table where the form is not found of any merger of the rows and columns in the row position merge 1/column 1. This research provides techniques to enrich the instance of a database, and with the use of illustrations, and then an approach to do the extraction of tabular HTML format can be done in a semi-automatic. In addition to that property in the table which is extracted can be distinguished from the contents of the cell which is a data table.

Keywords: HTML, Property Name of Table, Table Extraction, Website

1. INTRODUCTION

The data sources available on the Internet in various forms and formats, one of which is in the form of a table. The table consists of a cell, where each cell can contain a label/name attributes of the cell and cell data/content/attribute value[10]. For example that it find on the web site of a travel agent, it is showing data on the sales of airline tickets in the form of a table.

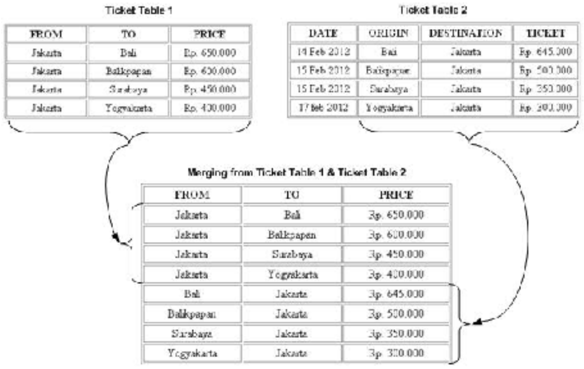
Data retrieval in the form of tables on the Internet performed to process data to the further process or does the merging of data extraction results with existing data.

Actually data retrieval on the Internet can be done by means of copy and paste, but this way is not easy if conducted at many tables. Table extraction approach is useful if you want to take a few tables from various sources on the Internet. The Illustrations can be seen in Figure 1, where the results of the data retrieval process of merging can be performed for further interoperability process.

Figure 1 shows two forms of tables about the ticket pricing information with the property name which is different but they have the same meaning. The table extraction will be useful to combine contents of both tables into one table only.

*Email Address: detty@staff.gunadarma.ac.id / dettydepe@yahoo.com

1936-6612/2011/4/400/008 doi:10.1166/asl.2011.1261



Ticket Table 1

FROM	TO	PRICE
Jakarta	Bali	Rp. 650.000
Jakarta	Balikpapan	Rp. 600.000
Jakarta	Surabaya	Rp. 450.000
Jakarta	Yogyakarta	Rp. 400.000

Ticket Table 2

DATE	ORIGIN	DESTINATION	TICKET
14 Feb 2012	Bali	Jakarta	Rp. 645.000
15 Feb 2012	Balikpapan	Jakarta	Rp. 500.000
15 Feb 2012	Surabaya	Jakarta	Rp. 350.000
17 feb 2012	Yogyakarta	Jakarta	Rp. 300.000

Merging from Ticket Table 1 & Ticket Table 2

FROM	TO	PRICE
Jakarta	Bali	Rp. 650.000
Jakarta	Balikpapan	Rp. 600.000
Jakarta	Surabaya	Rp. 450.000
Jakarta	Yogyakarta	Rp. 400.000
Bali	Jakarta	Rp. 645.000
Balikpapan	Jakarta	Rp. 500.000
Surabaya	Jakarta	Rp. 350.000
Yogyakarta	Jakarta	Rp. 300.000

Figure 1 Merge Illustration of Tickets Table

This research develops an approach to perform the extraction tables which are sourced from the Internet with HTML formatting. Table extraction performed on a simple table that are not joining the first row and first column, as well as to distinguish the contents of the cell as a property and the contents of cell which is the data. This is done by creating a database Property Name on the pre- processing stage. In addition there are 3 (three) techniques that are guaranteed to enrich the database instance Property Name in pre-l processing stage, namely: (i) do the words translation with online Dictionary, (ii). Looking for the similarity of the meaning from its word, (iii). Search for abbreviations. Tables that are present on the Internet and database Properties Name in the language of Indonesia is prepared.

2. STATE OF THE ART

A table is showing the data structured and the information related in the form of two dimensions[5]. The content of the table is the data presented in a cell. The table consists of a few boxes or cell, with the terms between the cell can form relationships columns and rows which have a attached meaning[7].

HTML and XML is the language used for representing semi structured data. HTML only pay attention view on the screen and not be used as an application storage / database, while XML not only determine the appearance on the screen, but XML can also be regarded as a database and can be processed directly by the application of other database or with query language[1].

For HTML documents is not formed by a fixed structure/clear, then it can be referred to as semi structure or unstructured document[5]. Extraction or can be referred to as the wrapper is part of the application that makes the source on the web that can be the source which can be queried if the database forms, and the source is in the form of semi structured[3].

Good HTML has a tag <head> and <body>, <head> define the header of HTML documents that contains information about HTML documents. Tag <head> not having the attributes but having special container in header <base>, <meta>, dan <title>. Tag <body> defines the beginning of the website content and covered with <body>…</body>. Tag <body> contains contents document that would appear in browser[8].

Approach to automated table extraction using a special characteristic of semi-structured format in an HTML table. Extraction tables conducted to distinguish 2 types of cell and cell identification label (property) and a data cell (instance). Property which is distinguished properties arranged on a table row and property on table column. Algorithm developed applicable for sample tables used in the research (not for other forms)[10].

Creating tables for the website with HTML begins with tags <table> and ends with the tag </ table>, tag <tr> and the end tag </tr> function to declare a row in the table, data tables have <td> start tag and the end tag </ td> serves to express the data / contents of the table, and table header, it has beginning tag <th> the end tag </th>[6.]

Approach to automated table extraction using a special characteristic of semi-structured format in an HTML table. Extraction tables conducted to distinguish 2 types of cell and cell identification label (property) and a data cell (instance). Property which is distinguished properties arranged on a table row and property on table column. Algorithm developed applicable for sample tables used in the research (not for other forms)[10].

An extraction table is done by creating a tree that is named Content Tree (CT) which is intended hierarchy. Firstly, it determined the hierarchy of connectedness of the content / data elements in the table. The second, strategy conducted during content creation tree is to map the tables in tabular form pseudo (pseudo table), then from the pseudo-table created a tree[4].

Document Object Model (DOM) is the base and the stand-alone language used to represent and make the connection between objects of different documents into HTML web pages or XML, with the tree structure as illustrative form (referred to as node tree)[2].

Xtractor algorithm also perform extraction on HTML tables and table data extraction stage consists the steps, namely[2]: i). Determine the initial HTML tag as parent, ii). Elaborating HTML tags are successfully extracted in order <tag> and </ tag> therein can be removed, iii). Specify a condition to stop, iv). An iterative process is performed to obtain the node, parent, child, sibling and their leaf.

3. METHODOLOGY

Extraction in HTML format table is done by taking into account the existing <table> tag on HTML tags and perform content matching cell with the property name database to determine the properties of the table position.

Based on a survey conducted on 100 tables that exist in the Indonesian language website, found that as many as 69 tables (69%) displayed a table with 1 row on the property, as well as 31 tables (31%) displayed on the Internet is a table with the most rows on the property, and no merging lines (join row) as well as the incorporation of columns (join column) at the property line. So in this study used a table is a table with the position of the property on line 1 above, except that the algorithm can also be run if found on the property table position in the leftmost column.

A. *Pre Processing on Property Name Database*

In the pre-processing stage, conducted a survey on 150 existing table forms on the Internet to collect property names used in the 150 table in Indonesian language. The survey results are stored in a property name database, and then enrich an instance of the property name database by see matching/similarity or likeness of the meaning of the word property name, as well as the usual abbreviations written on behalf of the property. The property name database on this study is in Indonesian language.

Table 1 is a property name that is on the property name database.

Tabel 1. Property Name Database

No	Property Name	No	Property Name
1	*Alamat* (address)	6	*Jenis Layanan* (service type)
2	*Alias* (alias)	7	Jumlah Penduduk (population)
3	*Batas Wilayah* (boundaries)	8	*Karyawan* (employee)
4	*Berangkat* (leave)	9	*Laki-Laki* (male)
5	*Class*	10	*Limit Harian Nominal* (Nominal Daily Limit)

The first way can be seen in Figure 2 that the way to enrich the instance in the process of translation is done by using an online dictionary. This method is done on the property name translation from Indonesian to English or vice versa, i.e. property name "*alamat*" (address), enriched by address. Enriching instance of the property name database can be done by three ways as shown in Figure 2.

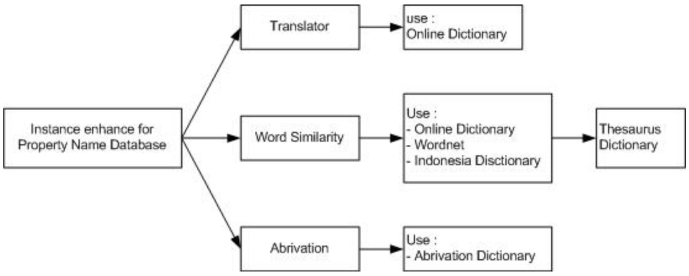

Instance enhance for Property Name Database
Translator
use : Online Dictionary
Word Similarity
Use : - Online Dictionary - Wordnet - Indonesia Disctionary
Thesaurus Dictionary
Abrivation
Use : - Abrivation Dictionary

Figure 2. Enriching Database Instance Property Names

B. *Thresholds Value for Table Property Detection*

Survey conducted on 150 of tables form available on the Internet to find the name of the property that is saved to the name property database on the research of table extraction pre-processing phase. Then are used 50 forms tables also taken from the Internet and being checked the 50 form properties names of each table to the name property database of the 150 table form. Results of a survey conducted get the thresholds value (TH) is 0.77. Formally the thresholds value obtained by using the following formula F.1:

$$\text{Threshold (TH)} = \frac{\sum_{i=1}^{T}(NP/KP)}{T} \qquad \text{F.1}$$

Here is the definition for the thresholds variable:

1. Thresholds (TH) are the limit value to determine a set of properties.
2. KP is the Number of Property on i^{th} table or the number of columns (1^{st} ... n^{th}) in i^{th} the table.
3. NP is the number of property name in the i^{th} table found in the property name database (RP).
4. T is the number of tables surveyed.

Number of property names found in the search results of property name (RP) database is formulated in F.2 as follows:

NP → KP (cell $_{1..n}$) ≈ RP, where NP Є P (Property) F.2

C. *Utilization Stages of Property Name Database*

Steps to check the property which done because of there were no *colspan* and no *rowspan* in <tr> ... </ tr> to-1 tag are as follows: i). Check the contents of cell (to form a word or string of words) taken from the line of unity (the tag <tr> ... </ tr> to-1), starting from the 1st column up to nth. Then fill the cell compared to a database name property (RP), ii). If at step one found any word similarity between the contents of the cell with the property name (RP) database, it is said that the contents of the cell is a member of the set of properties (P Є NP). Then repeat step 1 to the contents of cell in row 1 and the next column. (next column), iii). Having completed all cell contents in check, count the number of cell contents on the first row found in the property name (NP) database divided by the number of columns in the first row of the table (n), if the value is greater or equal to the threshold value (NP / n> = TH), if "conditions are met" it is said that the first row is a property that is on the row of the table, iv). If the value of NP / n is less than the threshold value (NP / n <TH), so do check for cell content from the first column of the table and began to line 1 to the to-q, is found in the database name property, v). If in step 4 found the word similarities between the contents of the cell with the property name (RP) database, it is said that the contents of the cell is a member of the set of properties (P Є NP). Then repeat step 4 for the contents of cell in column 1 of the next row. (next row)

After the contents of cell in column 1 line 1 until q finishes in check, calculate the distribution of cell contents were found (NP) with a number of existing row in column 1 of the table (q), if the result is greater than or equal distribution with threshold (NP / q> = TH), it is said that the first column cell content is property that is in the column of the table. But if there were no equality of outcome comparisons with cell contents of the database name property (NP / q <TH), then the table is the table that does not have the property.

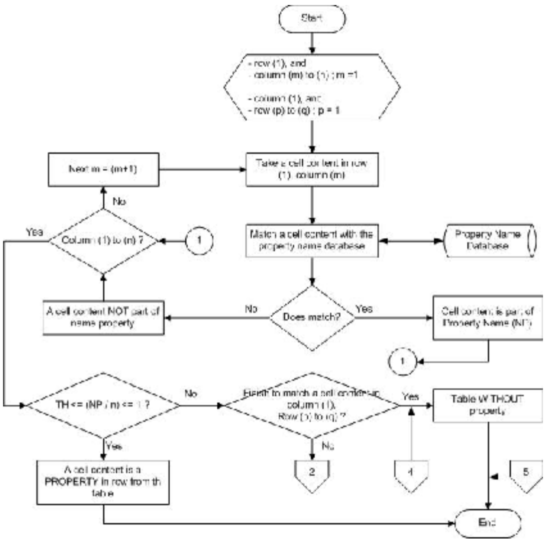

Start
- row (1), and
- column (m) to (n) ; m =1
- column (1), and
- row (p) to (q) ; p = 1
Take a cell content in row (1), column (m)
Next m = (m+1)
No
Yes
Column (1) to (n) ?
1
Match a cell content with the property name database
Property Name Database
A cell content NOT part of name property
No
Does match?
Yes
Cell content is part of Property Name (NP)
1
TH <= (NP / n) <= 1 ?
No
Finish to match a cell content in column (1), Row (p) to (q) ?
Yes
Table WITHOUT property
Yes
No
A cell content is a PROPERTY in row from th table
2
4
5
End

Figure 3. Stages Extraction Tables with Cell Contents Checking in Row 1 with Database Name Property

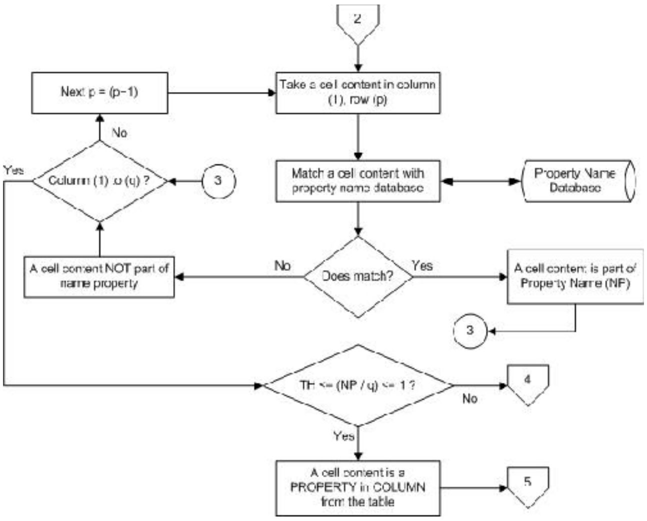

2
Next p = (p-1)
Take a cell content in column (1), row (p)
No
Yes
Column (1) to (q) ?
3
Match a cell content with property name database
Property Name Database
A cell content NOT part of name property
No
Does match?
Yes
A cell content is part of Property Name (NP)
3
TH <= (NP / q) <= 1 ?
4
No
Yes
A cell content is a PROPERTY in COLUMN from the table
5

Figure 4. Phases of Tables Extraction with Testing of Cell Content in first Column with Property Name Database

In Figure 3 shows the flow performed at the time of checking the contents of cell in row 1 and column 1 to n to find whether the property is in a table row. Then in Figure 4 is a continuation of the steps for checking the contents of the cell with the database name of the property on the 1st column starting from first row to the qth to determine if a table has a property on the column or table that does not have the property.

Algorithm 1 to check whether the property of the table in the position of row 1 or there on the 1st column or table that does not have the properties.

Algorithm 1. Find Property with Comparing Cell Content

```
m = 0
n = 0
read tag <tr>
If read tag <tr> then m = m + 1
      JumBaris = m
Loop;
If read tag <td> then n = n + 1
      JumKolom = n
Loop;
m = 1; n = 1; NP = 0 ; TH = 0,77 ;
n = read tag <td> in tag <tr>…</tr> ke-1
For n = 1 to JumKolom do
      isiCell (1,n) = value in tag <td>…</td>
      Function (WordSimilarityBaris) ;
Next n;
If (NP/JumKolom) > = TH then Table has Property in Row
Else
NP = 0 ;
m = read tag <tr>…</tr>
For m = 1 to JumBaris do
      Read tag <td>…</td> ke-1
      isiCell (m, 1) = value in tag <td>…</td>
      Function (WordSimilarityKolom);
Next m;
If (NP/JumBaris) >= TH then Table has Property in Column
Else Table hasn't Property
```

Then do a comparison to find the contents of cell in row 1 of the table with the database name property using the algorithm 2 and comparison algorithms to search the contents of cell in column 1 of the table with the property name database using the algorithm 3.

Algorithm 2. Word Similarity for Checking Row (1)

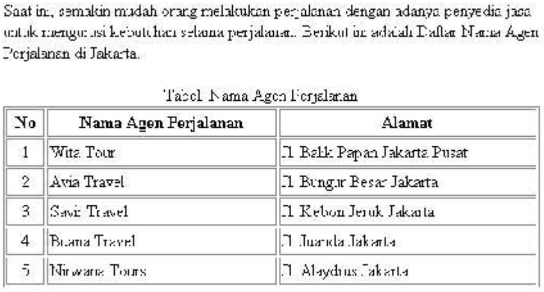
```
For r = 1 to jum_rec_database
read record (r) from Property name Database
If isiCell (1,n) = record (r) then NP = NP + 1
else
Next r;
```

Algoritma 3. Word Similarity for Checking Column (1)

```
For r = 1 to jum_rec_database
read record (r) from Property name Database
If isiCell (m,1) = record (r) then NP = NP + 1
else
Next r;
```

4. ILUSTRATION

Figure 5 is an example of a list of information about the travel agency (in Indonesian, because this research used object in Indonesian) that is presented in tabular form.



Saat ini, semakin mudah orang melakukan perjalanan dengan adanya penyedia jasa untuk mengurusi kebutuhan selama perjalanan. Berikut ini adalah Daftar Nama Agen Perjalanan di Jakarta.

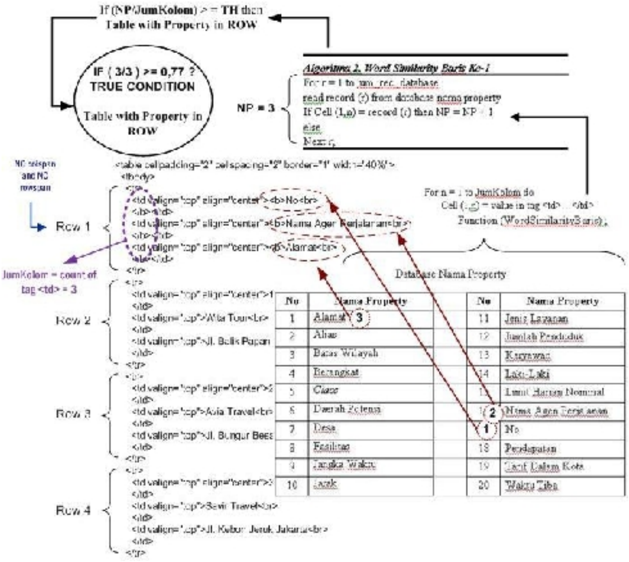
Tabel. Nama Agen Perjalanan

No	Nama Agen Perjalanan	Alamat
1	Wita Tour	Jl. Balik Papan Jakarta Pusat
2	Avia Travel	Jl. Bungur Besar Jakarta
3	Savir Travel	Jl. Kebon Jeruk Jakarta
4	Buana Travel	Jl. Juanda Jakarta
5	Nirwana Tours	Jl. Alaydrus Jakarta

Figure 5. Content Cell Comparison Table Illustration by Property Name Database

Figure 6 shows an illustration of algorithm passes. Number of columns is done by counting the number <td> tag that is in <tr> tag ... </ tr> to-1, and in the illustrations known <td> tag is 3 (three), so the variable *JumKolom*= 3.

Taken value inside the tag <td> ... </ td> on tag <tr> ...</ tr> to-1, then by using the algorithm 2, the name of an existing property in the Property Name Database matched. In the illustration, obtain a suitable number is 3. (NP = 3). Once the algorithm 2 is terminated, then back to the algorithm 1 with the condition: "IF (NP / JumKolom)> = TH then Table with property in ROW". Then calculate the number of names found property (NP) divided by the value contained in the variable JumKolom, and it is known that the threshold value used in this study was 0.77 which is the result of a survey of pre-processing stage. Provided that the condition is TRUE (3/3> = 0.77), so that in the illustration table is a table that has the property that the property is in position 1 is the top row.


If (NP/JumKolom) > = TH then
Table with Property in ROW
IF (3/3) >= 0,77 ?
TRUE CONDITION
Table with Property in ROW
NP = 3
NC colspan and NC rowspan
Row 1
Row 2
Row 3
Row 4
JumKolom = count of tag <td> = 3
Database Nama Property
3
2
1

```
Algoritma 2. Word Similarity Baris Ke-1
For r = 1 to jum_rec_database
read record (r) from database nama property
If Cell (1,n) = record (r) then NP = NP + 1
else
Next r,
```

```
For n = 1 to JumKolom do
  Cell (1,n) = value in tag <td> ... </td>
    Function (WordSimilarityBaris);
```

```
<table cellpadding="2" cellspacing="2" border="1" width="40%">
 <tbody>
  <tr>
   <td valign="top" align="center"><b>No<br>
   </b></td>
   <td valign="top" align="center"><b>Nama Agen Perjalanan<br>
   </b></td>
   <td valign="top" align="center"><b>Alamat<br>
   </b></td>
  </tr>
  <tr>
   <td valign="top" align="center">1
   </td>
   <td valign="top">Wita Tour<br>
   </td>
   <td valign="top">Jl. Balik Papan
   </td>
  </tr>
  <tr>
   <td valign="top" align="center">2
   </td>
   <td valign="top">Asia Travel<br>
   </td>
   <td valign="top">Jl. Bungur Bes
   </td>
  </tr>
  <tr>
   <td valign="top" align="center">3
   </td>
   <td valign="top">Savir Travel<br>
   </td>
   <td valign="top">Jl. Kebun Jeruk Jakarta<br>
   </td>
  </tr>
```

No	Nama Property	No	Nama Property
1	Alamat	11	Jenis Layanan
2	Alias	12	Jumlah Penduduk
3	Batas Wilayah	13	Karyawan
4	Berangkat	14	Laki-Laki
5	Cuaca	15	Limit Harian Nominal
6	Daerah Potensi	16	Nama Agen Perjalanan
7	Desa	17	No
8	Fasilitas	18	Pendapatan
9	Jangka Waktu	19	Tarif Dalam Kota
10	Jarak	20	Waktu Tiba

Figure 6. Illustration algorithm on HTML Source Code

5. CONCLUSSION

Merging rows or columns in the HTML tag marked with the tag colspan and tag rowspan. If the tag colspan and tag rowspan are not found in the first row of the table, the approach in this study can be used. Extraction table HTML format with this approach find property on the table that can be distinguished from the cell contents is the data table. In the illustration done in this study, the approach of table extraction can find out a table that has a property on the row position or find out the table with a property on the column position or find out the table that do not have a property.

This study was conducted to chart in a simple form, which did not reveal any table rows and columns merging both on property and in the data table, so for further research on the extraction of HTML table that will be used are merging table rows or columns on the content of a cell the data from the table.

ACKNOWLEDGMENTS

This work was supported by Study Center of Technology and Information System Gunadarma University Jakarta Indonesia.

REFERENCES

[1] Baumgartner, Robert; Flesca, Sergio; Gottlob, Georg. 2001. Visual Web Information Extraction with Lixto. Proceedings of the 27th International Conference on Very Large Data Bases. Page 119 – 128

[2] Gultom, Rudi; Sari, Riri Fitri; Budiardjo, Bagio. 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, Kristina ; Knoblock, Craig ; Steven, Minton. 2001. Automatic Data Extraction from Lists and Tables in Web Sources. Proceedings of Automatic Text Extraction and Mining Workshop (ATEM-01)

[4] Lim, Seung-Jin; Ng, Yiu-Kai. 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] Liu, Ying; Mitra, Prasenjit; Giles, C. Lee. 2008. A Fast Preprocessing Method for Table Boundary Detection : Narrowing Down the Sparse Lines using Solely Coordinate Information. The Eighth IAPR International Workshop on Document Analysis Systems, DAS '08. Page 431 – 438

[6] Ramadhan, R. 2008. Pembuatan Tabel (HTML 5). http://rizkyramadhansttg.wordpress.com/2008/07/15/pembuatan-tabel-html-5/, access date Sept 20, 2011

[7] Ramel, J.Y.; Crucianu, M.; Vincent, N.; Faure, C. 2003. Detection, Extraction and Representation of Tables. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)

[8] Ronggobramantyo, D. 2007. Belajar HTML yang merupakan Dasar dari Pembuatan Website. http://www.dhimasronggobramantyo.com/artikel/Belajar_HTML_yang_merupakan_dasar_dari_pembuatan_website, access date Sept 20, 2011

[9] Sridhara, Giriprasad ; Hill, Emily ; Pollock, Lori ; Shanker, K. Vijay. 2008. Identifying Word Relations in Software: A Comparative Study of Semantic Similarity Tools. Proceedings of the 2008 The 16th IEEE International Conference on Program Comprehension. Page 123-132

[10] Tengli, Ashwin; Yang, Yiming; Ma, Nian Li. 2004. Learning Table Extraction from Examples. Proceeding COLING '04 Proceedings of The 20th International Conference on Computational Linguistics