Issues in Information Extraction from Web-Tables

Mahesh A. Sale
M.Tech. Computer, Department of Computer Technology, Veermata Jijabai Technological Institute, Mumbai-19, Maharashtra, India

Pramila M. Chawan
Associate Professor, Department of Computer Technology, Veermata Jijabai Technological Institute, Mumbai-19, Maharashtra, India

Prithviraj M. Chauhan
Project Manager, Morning Star India Pvt. Ltd., Navi Mumbai, Maharashtra, India

Abstract

Automating the process of table search and extraction is a challenging problem for several reasons. It is about information extraction from HTML tables published inside web pages. Web tables present challenges to information retrieval because of different layouts, diverse media, different cell types and table elements, etc. Before extracting the information from web tables, one needs to analyze the problems associated with the same. In this paper, we identified the issues related to information extraction from web tables, in which we analyzed the web tables with respect to its different components, properties, layouts etc. We also identified some problems with table extraction and suggested some solutions for the same.

1. Introduction

Web Tables are an important way of publishing information to the web user. The wide spread use of web tables on web is for the reason that they show relational data concise manner and thus easy to read by human beings. Automation of table extraction process has applications web mining, information retrieval, knowledge acquisition, and summarization. A table is two-dimensional grid of cells. Each cell either represents a label (attribute) or a data (attribute value). The table extraction process involves identifying these two types of cells and establishing a semantic relationship between the two.

Automating the process of extracting the information from web tables is a challenging problem web tables are designed for human beings in order to increase the ease of understanding the information and not for the computers. Thus a web table unambiguous for a human may not be that much easy for a computer to parse and identify the information in it.

2. Web Table Analysis

2.1. HTML Table

In HTML the table contents are confined to the body of text within <table> and </table> html tags. These HTML tags specify how the table is rendered including structure and visual appearance of the table. The structure of the table is associated with the rows and columns in the table and how the text is stored in each cell of the table. Before extracting the information from web tables one should note that, in HTML, rows are first created and then the data is inserted into the cells. Thus a web table analysis should begin with rows and then columns. The visual appearance of table refers to the rendering of text within the cells, for example font size, font style, colour, text alignment, background colour etc.

2.2. Attribute-Value pairs

The information in web tables is extracted in terms of attribute-value pairs. An attribute constitute one or more labels where a label is formed by a group of words. On the other hand a value refers to quantities of attributes in their respective units like number, currency, degree, percentage, salary, time, date etc. Thus, before extracting the information from tables, one has to recognize these units while scanning the table. Also, some special character strings such NIL, N/A and ranges of quantities (e.g. 30,000-50,000) are needs to be recognized. It has been seen that a value associated with an attribute rarely has more than one quantity.

2.3. Heading and Content rows

The rows of a table can be categorized as heading rows and content rows. A heading row is defined as a row which contains the attribute labels of the table. Usually there is only a single heading row which is located in top most rows of the table. But there can be more than one heading rows in case of Composite Table described in Section 4.2. The content row contains the values of attributes located in heading row. The heading row contains significantly less number of cells in a row. Also, it does not label-quantity pairs and the visual appearance of heading row is different from content row.

2.4. Row and Column spanning

In HTML tables, row and column spanning is often used to increase the width of a cell beyond its default width. The value of rowspan or colspan attributes in HTML is always a positive integer. For example, “rowspan=3” means the cell spans three rows consecutively, starting from the current row. To deal with row spanning, the simplest way to duplicate the label in each row it spans and setting the value of rowspan attribute to 1. The same thing applies to colspan also.

3. Classification of Web-Tables

In this section, we classified web tables into three categories according to the objective of the <table> tag as follows:

a) Tables for layout of Web page

The tables for web layout are used to tune up the layout patterns of web page.

b) Tables for emphasis

Tables for emphasis are used for emphasizing a list of links, pages etc. It does not include any logical or informative contents of table.

c) Genuine Tables

Genuine tables are the two-dimensional structures where one can establish the relationship between attributes and their values [10]. Genuine tables have the following characteristics:

1) The value of BORDER attribute of <table> tag can be more than one

2) The table contains more than two cells.

For extracting the exact information from web tables, one has concentrate on genuine tables only, because they include knowledge and data of interest and thus regarded as sort of database.

4. Formats of Web-tables

In order to extract the information from web tables one has to be aware of different possible formats of the tables. Web tables can be in different sizes, shapes or in the formats suitable for their underlying information. Tables can be considered as *primitive table* or *composite table*.

4.1. Primitive Table

A primitive table contains attributes and their corresponding values arranged as a grid of cells.

The following aspects define the format of the primitive table. Each of these aspects has its own way to present and interpret the underlying information.

a) Spanning Aspects

The formats of most of the tables are defined by the spanning aspects, where the table is shaped by spanning cells or super rows. Spanning cells spans two or more than two cells in column or row manner. A super row accommodates all the cells in a row, in other words super row is similar to spanning cells in an entire row.

b) Dimension Aspects

A web table can be either one-dimensional or two-dimensional. In on one-dimensional table the attribute and value cells are arranged in horizontal or vertical manner. Here, each attribute is followed by the associated values in column-wise or row-wise style. In a two-dimensional table the value cells are represented in perpendicular manner [9]. If the attribute labels are arranged row-wise, each column represents a set of values corresponding to that column and if column-wise, each row represents the same.

4.2. Composite Table

A Composite table accommodates two or more primitive tables. The following two aspects define the format of the composite table:

a) Corresponding Aspects

The different primitive tables in a composite table similar or dissimilar in order to compare, summarize and understand individual piece of information by human. When two tables have same attributes they are said to be the similar primitive tables. Thus the information contained in them should be treated similarly. On the other hand, when two tables have different set of attributes the primitive tables are said to be dissimilar [9]. These tables represent different but related information which should be interpreted.

b) Combining Aspects

When forming a composite table, primitive tables are combined in horizontal or vertical manner. For similar primitive tables both horizontal and vertical manners are followed and for dissimilar primitive tables vertical style is followed.

5. Table Metadata

There are no formal standards or rules exist for designing table. In order to facilitate the table information processing and extracting an extensive and universal metadata specification is needed [10]. Also a standard representation can highly reduce the efforts of information extraction from web tables.

The table metadata can be classified in following categories:

a) Environment

The metadata about table environment includes the information about the web page where the table is located. For e.g. Document Title, Document Author, Document Source or Origin, Document creation date, the X-Y position of the table in document. This metadata can ease the table searching process if users don’t the details of the table contents.

b) Affiliations

The Table Affiliated Metadata includes different affiliated elements like Table Caption, Footnotes, and References etc. Table Caption describes the table which appears along with the table, above or below it. For e.g. “Table 1. Executive Details”. Table Footnote explains the information in the table, which appears below the table. Table Reference discusses the contents of the table.

c) Frames

The Table Frame Metadata indicates whether there are frames around a table, with values right, left, top, bottom, all and none.

d) Layout

The Table Layout Metadata indicates the visual appearance of the table, including Number of columns, Number of Rows, Row headers, Column headers, Table width, Table length, Column width, Row length, Horizontal alignment.

e) Cell Type

The Cell Type Metadata indicates the type of the contents of a cell in table and thus can be useful to determine the table type. The cell contents may be numeric, textual, symbolic, formula, mathematical equations, images or even another table.

f) Cell Contents

The Table Cell Contents data includes the values in the cells of the table. These values enable users to search the tables on the basis of the cell contents of the table.

6. Problems

In this section we presented the problems encountered in extracting the information from web tables. Also, we suggested some possible solutions on these problems. The problems are categorized in two parts – Table Location problems and Table Extraction problems.

6.1. HTML Tables – Location Problems

Locating the table of interest on a web page is trivial task for a human being. However, algorithmically finding the table of interest on a web page is non-trivial, even though the system can tell that the web page is of interest to the application. Although there are many challenges, we identified some challenges of table location as follows:

a) HTML lists instead of <table> tags

Sometimes the data of interest is organized as a single-column table, where the header of the column indicates the attribute associated with the contents of the column. But the table is not included under <table> tag, but rather with a <ul> tag, treating it as an HTML list. In such cases, the object to which the attribute is associated is described in the lines or paragraphs located above the list.

b) Superfluous information included in table rows

Some of the data included under <table> tag may not be of interest or it may be an invalid data. Normally such data is the graphical buttons on the web page, summary of the information displayed in table or few sentences regarding the object for the table is displayed.

c) Piecemeal Tables

Some of the tables display only a limited number of rows by default. For the rest of the rows we have click on the provided links like “See more rows”. This can be seen in case of tables with large number of rows. Algorithmically, one has to simulate the click operation by heading towards the rows pointed by the link.

d) Information located on different pages

It may be the case that, the information of interest to be extracted from web tables may be located in different tables on different web pages. In short, all the attributes and their values may not be located in a single table. Also, the information may be distributed on the same page or different page, but in both cases one has to identify the targeted attribute-value pairs.

e) Multiple HTML frames

There may be multiple HTML frames (also called as panes) on the web page. One has to identify the frame which of interest.

f) Tables in plain text ASCII format

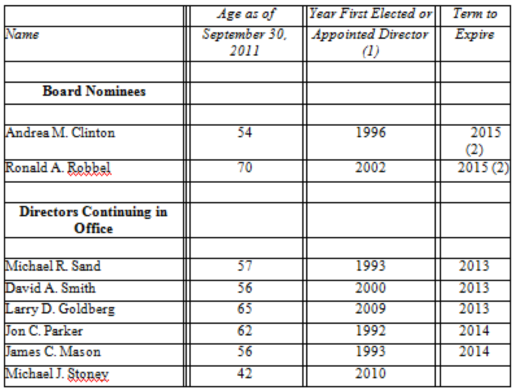
Some websites (like [1]) gives a link to the table of interest, which points to a table in text format. In HTML, a table is demarcated by a pair of <table>...</table> tags. Thus it is trivial to interpret the table in a web page by identifying the data contained within these two tags. The rest of the table contents can then be identified by similarly identifying the <tr> and <td> tags which represents a row and a cell in HTML table respectively.

But this approach is not suitable to identify the information in plain text ASCII tables. Such tables can be visually interpreted by human eye. But algorithmically it is much more difficult to interpret it. Unlike the HTML tables, the data in plain text ASCII tables is structurally organized by using spaces and tabs [6]. Thus the challenge here is to interpret structural information of the table from spaces, tabs and ASCII character sequences like “#”, “-” etc. for creating the borders of the table.

6.2. HTML Tables – Extraction Problems

Not only is it very easy for a human being to locate the table of interest on a web page, but also its easy for a human to parse the table and determine the attribute-value pairs, irrespective of the view of the table. It very straightforward for a human to semantically match the table contents with the target database schema. But constraints are needed to be imposed when we algorithmically match a source table with respect to a fixed target view. Algorithmically finding these semantic matches is significantly harder. We investigated the following challenges that can occur during information extraction:

Table 1. Executive details table [1]

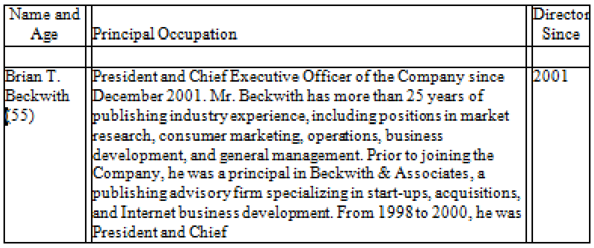


	Age as of	Year First Elected or	Term to
Name	September 30, 2011	Appointed Director (1)	Expire
Board Nominees			
Andrea M. Clinton	54	1996	2015 (2)
Ronald A. Robbel	70	2002	2015 (2)
Directors Continuing in Office			
Michael R. Sand	57	1993	2013
David A. Smith	56	2000	2013
Larry D. Goldberg	65	2009	2013
Jon C. Parker	62	1992	2014
James C. Mason	56	1993	2014
Michael J. Stoney	42	2010	

a) Genuine Tables

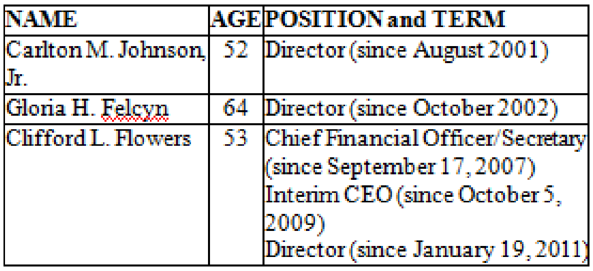
After locating the table in a web page, the next important task is to identify the genuine tables. The genuine tables are those which contain the valid information. A web page may contain many tables and every table may not be of interest to us. For identifying a genuine table one should consider the things like attributes, their values, number of rows and columns, table captions etc.

Table 2. Executive details table [2]



Name and Age	Principal Occupation	Director Since
Brian T. Beckwith (55)	President and Chief Executive Officer of the Company since December 2001. Mr. Beckwith has more than 25 years of publishing industry experience, including positions in market research, consumer marketing, operations, business development, and general management. Prior to joining the Company, he was a principal in Beckwith & Associates, a publishing advisory firm specializing in start-ups, acquisitions, and Internet business development. From 1998 to 2000, he was President and Chief	2001

Table 3. Executive details table [3]



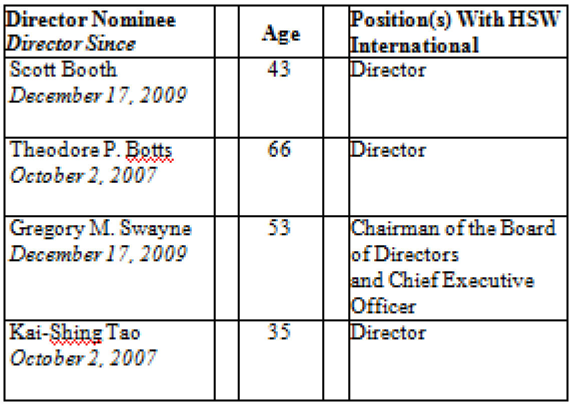
NAME	AGE	POSITION and TERM
Carlton M. Johnson, Jr.	52	Director (since August 2001)
Gloria H. Felcyn	64	Director (since October 2002)
Clifford L. Flowers	53	Chief Financial Officer/Secretary (since September 17, 2007) Interim CEO (since October 5, 2009) Director (since January 19, 2011)

b) Formatting differences and Lexical Variants

One of the challenging issues in information extraction is to deal with formatting difference and lexical variants in the tables whose semantic meaning is same. The tables in the filings submitted by various companies to SEC (U.S. Securities Exchange Commission) are good examples of this. This concept is often referred as CDS (Common Data Set). CDS is an initiative to define a standard format for the filings submitted to SEC. However the tables that appear on the filings submitted to SEC differs greatly in terms of formatting, lexical variations in labels, missing or extra portions in those tables, etc. The process of mapping the tables in various formats to the unified predefined structure is a very complex task.

For example, Table 1, Table 2, Table 3 and Table 4 are the formats of Executive Details tables in the filings submitted to SEC. The unified structure of the table expects the attributes as Director Name, Age, Position and Director since. But these attributes are represented using different labels as shown in above mentioned tables.

Table 4. Executive details table [4]



Director Nominee ***Director Since***		**Age**		**Position(s) With HSW International**
Scott Booth *December 17, 2009*		43		Director
Theodore P. Botts *October 2, 2007*		66		Director
Gregory M. Swayne *December 17, 2009*		53		Chairman of the Board of Directors and Chief Executive Officer
Kai-Shing Tao *October 2, 2007*		35		Director

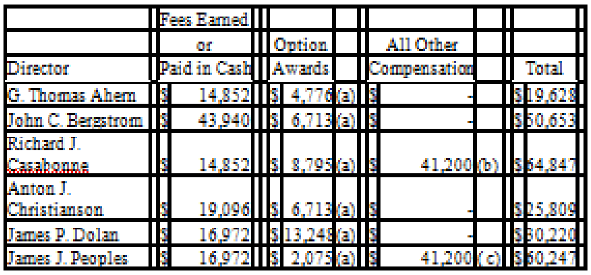
c) Merged Attributes or Values

The column labels represent the attribute associated with that column. Many of the times a single column accommodates more than two attributes and thus their corresponding values. As shown in Table 2 and Table 3 the attributes, Name-Age and Position-Term respectively, are given in a single column.

d) Attributes' position

In most of the cases the attribute labels are located in top most rows of the table. But there may be the case that they are along the column of the table or both. Such position of attributes decides the further procedure to be followed for extracting the cell information.

Table 5. Director compensation table [5]



	Fees Earned						
	or	Option		All Other			
Director	Paid in Cash	Awards		Compensation		Total	
G. Thomas Ahern	$ 14,852	$ 4,776	(a)	$ -		$ 19,628	
John C. Bergstrom	$ 43,940	$ 6,713	(a)	$ -		$ 50,653	
Richard J. Casabonne	$ 14,852	$ 8,795	(a)	$ 41,200	(b)	$ 64,847	
Anton J. Christianson	$ 19,096	$ 6,713	(a)	$ -		$ 25,809	
James P. Dolan	$ 16,972	$ 13,248	(a)	$ -		$ 30,220	
James J. Peoples	$ 16,972	$ 2,075	(a)	$ 41,200	(c)	$ 60,247	

e) Identifying the attributes from labels

Normally, we identify the attribute associated with a column by the label of the column (which is normally located at the rows or columns at top most positions). But sometimes the labels are not a single word entity; instead they are composed of one or more sentences. Thus, in order to identify to which attribute such label is associated, we need to separately parse the sentence and identify the attribute.

f) Super-Row Labels

A super-row is a row which spans the entire row of the table. The labels namely "Board Nominees" and "Directors Continuing in Office" correspond to super-row labels.

g) Synonyms

The labels like "Salary" and "Compensation" represent the different attributes but their meaning is same. One should consider such synonyms while identifying and extracting the attributes and values.

h) Unwanted information

The web table may contain some columns not of interest for target schema. For ex. some columns of the table may contain photos or links to other websites more details. This information should be regarded as extra information.

7. Conclusion

Identifying and extracting the information from web tables is algorithmically a challenging process. As apart from the standard syntax, there is no any standard format of presenting the data using tables in HTML, the complexity varies from source to source. We encountered a lot of issues of information extraction from web tables from different sources. We analyzed the web tables targeting the important issues of the tables like the attributes, values, heading rows content rows etc. For identifying the table of interest from web page, we classified the tables and defined some of the formats of the web tables. In addition, in order to ease the table identification process, we identified the possible metadata associated with a web table. The actual problems encountered in extracting information from web tables are explored. We believe that the issues and problems related with web table extraction will vary according to the source of the web.

8. References

[1] http://www.sec.gov/
[2] http://sec.gov/Archives/edgar/data/729156/000114420411 069351/v242565_def14a.htm
[3] http://sec.gov/Archives/edgar/data/ 836564/ 00010196871 1003893/patriot_def14a.htm
[4] http://sec.gov /Archives/edgar/ data/1368365/ 000 13 68 3 651100 0056/def_proxy.htm
[5] http://sec.gov/Archives /edgar/data/ 729156/ 000114420 41106 93 51/v242565_def14a.htm
[6] Ashwin Tengli, Yiming Yang and Nian Li ma, School of Computer Science, Carnegie Mellon University, Pittsburg, "Learning Table extraction from examples".
[7] Yingchen Yang, Wo-Shun Luk, School of Computing Science, Simon Fraser University Burnaby, Canada, "A Framework for Web Table Mining", ACM WIDM'02, November 8, 2002, McLean, Virginia, USA.
[8] Hidetaka MASUDA and Shuichi TSUKAMOTO, School of Engineering, Tokyo Denki University, Tokyo, Hiroshi NAKAGAWA, Information Technology Center, The University of Tokyo, "Recognition of HTML Table Structure".
[9] Nattapon Harnsamut, Naiyana Sahavechaphan, Large-scale Simulation Research Laboratory, National Electronics and Computer Technology Center, Pathumthani, Thailand, "Mining for Attributes and Values in Tables", ACM MEDES'10 October 26-29, 2010, Bangkok, Thailand.
[10] Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles, The College of Information Sciences and Technology, Pennsylvania State University, University Park, "TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries", ACM JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.