Extracting Table Information from the Web

Yeon-Seok Kim and Kyong-Ho Lee

Dept. Computer Science, Yonsei University,
134 Shinchon-dong, Seodaemun-ku, Seoul, 120-749, Korea
yskim@icl.yonsei.ac.kr, khlee@cs.yonsei.ac.kr

1 Introduction

With the ubiquity of the Web, the volume of Web documents continues to grow at a rapid speed. Since the Web is a vast source of information, extracting useful information from Web documents is important.

HTML (Hypertext Markup Language), which is a format for visual rendering of Web documents, defines tag `<TABLE>` for representation of a table. On the other hand, most of the existing HTML documents use `<TABLE>` tags to present a formatting layout of a document. As a prerequisite for information extraction from the Web, it is required to determine whether `<TABLE>` tags are used to present genuine tables or not.

Generally, a table is a facility for presenting relational information structurally and concisely. This paper defines a table as an array of relational data. Specifically, we regard a table that relates an attribute and its value, as a genuine table as reported in previous works. In this paper, set of attribute cells and set of value cells are defined as an attribute area and a value area, respectively.

Most previous works concerning table identification in HTML documents are based on a specific domain or take a lot of training data and time. This paper presents an efficient method for identifying tables in HTML documents prior to extracting information from the Web.

2 Table Detection

The proposed table detection method consists of two phases as shown in Fig. 1: pre-processing and attribute-value relations extraction. Additionally, the relations extraction consists of three steps: area segmentation, checking a syntactic coherency of a value area, and checking a semantic coherency between attribute and value areas.

2.1 Preprocessing

The proposed preprocessing detects tables based on general characteristics of genuine or non-genuine tables. For this purpose, eight rules were devised based on a careful examination of general characteristics of `<TABLE>` tags as shown in Fig. 2.

2.2 Relations Extraction

The relations extraction deals with the `<TABLE>` tags that are undetected at the pre-processing step. As mentioned before, the relations extraction method consists of

Fig. 1. The proposed table detection process.

Rule (1) - IF: A <CAPTION> tag exists.	THEN: The table is a genuine table.
Rule (2) - IF: The size of a table is 1×1 .	THEN: The table is a non-genuine table.
Rule (3) - IF: The table does not have any characters.	THEN: The table is a non-genuine table.
Rule (4) - IF: Most cells include only hyperlink.	THEN: The table is a non-genuine table.
Rule (5) - IF: Most cells consist of images.	THEN: The table is a non-genuine table.
Rule (6) - IF: Most cells are empty.	THEN: The table is a non-genuine table.
Rule (7) - IF: The table includes <th> tags AND Existence of <td> tags at the right or down sides.	THEN: The table is a genuine table.
Rule (8) - IF: A <TABLE> tag includes nested <TABLE> tags.	THEN: The table is a non-genuine table.

Fig. 2. The proposed preprocessing rules.

three parts: area segmentation, syntactic coherency checkup, and semantic coherency checkup. Detailed explanation of the proposed method is given below.

Area Segmentation. First of all, the proposed method segments a table area into attribute and value areas. If the size of a table is 1×2 (and $1 \times n$) or 2×1 (and $n \times 1$), the first column or row corresponds to an attribute area. If the size of a table is 2×2 and do not have any span, the first row (or column) would indicate an attribute area, there by making the second row (or column) a value area. For tables of size 1×2 , 2×1 , or 2×2 , the proposed method cannot check for syntactic coherency in a value area because there exists only one value for each attribute. The method checks for semantic coherency between attribute and value area. Specifically, for 2×2 tables with no span, existence of a semantic coherency should be checked by rows or columns. On the other hand, a table with the size of $1 \times n$ (or $n \times 1$), the method checks for syntactic coherency by rows (or columns). Tables with the size of $2 \times n$ or $n \times 2$ ($n \geq 2$) may not have any value area. For example, if the first row or column in a table consists of one

cell, the proposed method regards the table as a non-genuine table because no value area exists. The proposed method considers the first row with single cell and the next row as main attribute area and sub-attribute area, respectively. In the case of a $2 \times n$ table where the first row spans not whole columns but more than one column, then if the first row were an attribute area, the next row would be a sub-attribute. Thus, the proposed method regards the first column and the remaining columns as an attribute area and value areas, respectively. For the rest of $2 \times n$ (or $n \times 2$) tables, it is possible to apply the syntactic coherency checkup by rows (or columns) and the semantic coherency checkup in both directions. For example, if the proposed method regards the first column (or row) as an attribute area, the method can check for syntactic coherency and semantic coherency horizontally (or vertically). If the proposed method regards first row (or column) as an attribute area, then the method just examines for semantic coherency vertically (or horizontally). Furthermore, we segment tables, whose size is greater than or equal to 3×3 , into attribute and value areas like $2 \times n$ (and $n \times 2$) tables. Detailed description on the coherency checkup of segmented areas is as follows.

Syntactic Coherency Checkup. If a table has more than one value for each attribute row-wise (or column-wise), the proposed method examines the values existence of row-wise syntactic coherency (or column-wise). For this purpose, we define "row (or column)-wise coherency" as the average value of the coherency rows (or columns) that compose a value area as follows:

$$\text{Row (or column)-wise coherency} = \frac{\sum \text{row (or column) coherency}}{\text{The number of rows (or columns)}} \quad (1)$$

where the row (or column) coherency is computed by,

$$\text{Row (or column) coherency} = W_1 \times \text{data type coherency} + W_2 \times \text{length coherency}. \quad (2)$$

The data type coherency of Equation (2) is defined by,

$$\text{Data type coherency} = \frac{\text{The number of cells that share a major data type}}{\text{The total number of cells in a row (or a column)}} \quad (3)$$

where "major data type" means the data type with the highest frequency in the constituent cell of the corresponding row or column. The proposed method first sets the type of a data cell as image type or form type according to the kind of tags used in the cell. For the remaining cells, the proposed method determines their data types based on their contents. The proposed method defines 15 data types. Specifically, the textual patterns and keywords are used to recognize some of these data types.

The length coherency means how similar the lengths of the constituent cells are. It is denoted as the frequency of cells of which the length is within the range of ∞ , the average length of cells $\times 0.5 \leq \infty \leq$ the average length of cells $\times 1.5$, and is computed by,

$$\text{Length coherency} = \frac{\text{The number of cells with a length within the range } \alpha}{\text{The total number of cells in a row (or a column)}} \quad (4)$$

If it is possible to check coherency for a table in both horizontal and vertical direction, the method selects the larger value as the coherency value of the table. If the value of a table coherency is larger than a predefined threshold, $ThCoherency$, the

table is deemed to be a genuine table. Otherwise, the proposed method apply the semantic coherency checkup to the table.

Semantic Coherency Checkup. If a table is found to be inappropriate for the syntactic coherency checkup or if its table coherency value is smaller than *ThCoherency*, then the proposed method examines whether there exists any semantic match between an attribute area and its value area based on auxiliary information. The auxiliary information includes keywords or textual patterns, which correspond to values of specific attributes, as well as data types of value areas. The method regards `<TABLE>` tags with semantic relations between attribute and value areas as genuine tables. For instance, a table, in which an attribute 'e-mail' has the values with the pattern of "`\w+\@ \w+ \. \w+`", is regarded as a genuine table.

3 Experimental Results

To evaluate the performance of the proposed method, we experimented with 11,477 `<TABLE>` tags from 1,393 HTML documents. For comparison, we used the same data as in Wang and Hu's method [1]. The `<TABLE>` tags consist of 1,675 genuine tables and 9,802 non-genuine tables. The threshold of a table coherency, *ThCoherency*, the weight of data type coherency, and the weight of length coherency were experimentally determined as 0.54, 0.6 and 0.4, respectively. Experimental results show that the method has performed better compared with previous works, resulting in a precision of 97.54% and a recall of 99.22%. Particularly, unlike some conventional methods, this method does not need to be trained on a large volume of data in advance. Meanwhile, the proposed method could not detect non-genuine tables with a syntactic coherency or genuine tables with many empty cells.

Reference

1. Wang, Y., Hu, J.: Detecting Tables in HTML Documents. Proc. 5th IAPR Int'l Workshop on Document Analysis System (DAS'02), Princeton, USA (2002) 249~260