Expert Systems with Applications 42 (2015) 929–937

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Table understanding using a rule engine

Alexey O. Shigarov

*Institute for System Dynamics and Control Theory of SB RAS, Lermontov st. 134, Irkutsk 664033, Russia*

ARTICLE INFO

*Article history:*
Available online 7 September 2014

*Keywords:*
Table understanding
Table canonicalization
Information extraction from tables
Unstructured tabular data integration
Table model

ABSTRACT

The paper discusses issues on the conversion of tabular data from unstructured to structured form. Particularly, we propose an approach to table understanding (i.e. recovering semantic relationships in a table), which is designed for unstructured tabular data integration. Our approach is based on using a rule engine. It is assumed that spatial, style (typographical), and natural language information can be used for table analysis and interpretation. The *CELLS* system based on the approach has been developed for integrating unstructured tabular data presented in Excel spreadsheet format. Experimental results show that the approach and system can be applied to a wide range of tables from statistical and financial reports.



## 1. Introduction

Nowadays, many researchers in data management (e.g. Doan et al., 2009; Ferrucci & Lally, 2004; Inmon & Nesavich, 2007) note that issues on unstructured data management and integration become increasingly important. The term "unstructured information/data" usually refers to any information that does not have a predefined formal data model or does not fit into a table of a relational database. If unstructured information contains some text (e.g. plaintext, PDF, or Word documents) then it is called "unstructured textual information/data". More accurate terms "weakly structured" and "semi-structured documents" (Feldman & Sanger, 2006) are used to indicate unstructured textual information.

The documents may contain tables which do not have any formal data model. These tables are intended to be interpreted by humans but not designed for high-level machine processing like SQL queries. Therefore, in the sense defined above, these tables are examples of unstructured textual information. By analogy, they may be called "unstructured tabular information/data".

Automation for transforming tabular information into structured form has important applications in problems of data management, information extraction, and document analysis systems. There are the following problems which can be considered as the conversion of tabular information from unstructured to structured form.

- Table canonicalization (Douglas, Hurst, & Quinn, 1995; Tijerino, Embley, Lonsdale, Ding, & Nagy, 2005) is transformation of a table to the canonical form that fits into the table of relational database.
- Information extraction from tables (Embley, Hurst, Lopresti, & Nagy, 2006a) is analogous to the task of information extraction from texts and consists in extracting selectively facts to generate a target database.
- Table understanding (Embley et al., 2006a) consists in recovering relationships among data values, labels (attributes), and dimensions (domains). In general case, as Hurst (2001) notes, the table understanding involves the following steps: (1) table location (to detect positions of a table inside a source), (2) table recognition (to recover individual cells), (3) functional analysis (to find attributes and data in cells, i.e. to recover cell roles), (4) structural analysis (to recover relationships between cells), and (5) interpretation (to extract facts from a table).

The present work is restricted to the issues: how to recover relationships of table elements (i.e. cell-role, label-value, label-label, and label-dimension pairs). In terms of Hurst (2001), we propose to automate the following steps of table understanding: functional analysis, structural analysis, and interpretation of a table.

Our approach to table understanding is based on the use of a rule engine and table analysis rules. It is expected that facts which are used in the process of logical inference may include information about spatial, style (typographical) and natural language content of tables. The implementation of rule sets for different table forms provides the processing of a wide range of tables having complex structures. The *CELLS* system based on the proposed approach has been developed for integrating unstructured tabular data. It allows extracting data from tables presented in Excel spreadsheet files. The obtained experimental results demonstrate that the system can be applied to input data from tables into a database.

*E-mail address:* shigarov@icc.ru

http://dx.doi.org/10.1016/j.eswa.2014.08.045
0957-4174/

## 2. Related work

Depending on presentation level of a table, the table understanding requires to solve different tasks (steps), such as location, recognition, analysis, and interpretation of a table, in terms of Hurst (2001). Detailed surveys of methods and systems which are devoted to these problems can be found in the following papers (Embley et al., 2006a; Embley, Lopresti, & Nagy, 2006b; Lopresti & Nagy, 2000; e Silva, Jorge, & Torgo, 2006; Zanibbi, Blostein, & Cordy, 2004; Zanibbi, Blostein, & Cordy, 2008).

There is a huge amount of ways to portray a table. Table features originate from typographical standards, corporative practice, ad hoc software, data formats, and human inventiveness. It leads to the complexity of table understanding. The existing methods and systems related with the enumerated above steps of table understanding are based on different approaches, e.g. heuristic, machine learning, dynamic programming, or probabilistic methods. However, all of them use some assumptions about table structures to reduce the complexity of own tasks. Usually, those assumptions are embedded in their algorithms. It significantly constrains a range of tables that can be efficiently processed by these algorithms.

The current state of research in this area does not allow to say that the problems of table understanding are completely solved. The most studies devoted to the problems of low-level table processing, such as location and recognition of tables from document images and plain-text. Meanwhile, the issues of table understanding (including analysis and interpretation) remain less studied in the case of unstructured tabular information presented in high-level formats of a word processor or spreadsheet.

In the paper, we discuss only methods related with the steps of table analysis and interpretation. Particularly, the following papers (Douglas et al., 1995; Embley, Tao, & Liddle, 2005; Gatterbauer, Bohunsky, Herzog, Krpl, & Pollak, 2007; Hurst, 2000; Kim & Lee, 2008; Pivk et al., 2007; e Silva et al., 2006; Tijerino et al., 2005; Wang, Wang, Wang, & Zhu, 2012) made significant contribution to solving these problems of table understanding.

In the papers (Douglas et al., 1995; Tijerino et al., 2005) the approaches to table canonicalization are considered. The method for interpretation and canonicalization of tables which are contained in specifications used in construction industry is suggested by Douglas et al. (1995). It is based on natural language processing using domain ontology (i.e. a sub-language of construction industry specifications).

Another technique for table canonicalization proposed by Tijerino et al. (2005) is based on a library of frames containing knowledge about lexical content of tables. Each frame describes a data type using regular expressions, dictionaries, and open resources like the lexical database WordNet.[1] The frame is used to assign data types to table labels and values.

Embley et al. (2005) proposed methods for location of tables in HTML pages, and information extraction from them. It is assumed that a table may have nested tables on linked pages. In particular, in order to detect attributes (labels) and data values in cells they use ontologies developed specifically for information extraction. In addition to objects, relationships and constraints an extraction ontology includes a set of data frames which are associated with sets of objects. Those data frames allow binding table content with objects of the ontology using regular expressions. As well, in table analysis they use several table recognition heuristics on table structures and content.

Wang et al. (2012) consider the problem of understanding a web table as associating the table with semantic concepts presented in a knowledge base. In particular, they use Probase[2] as that knowledge base. This method can be applied only for HTML tables with a very simple structure without merged cells, when each row of a table, excluding a single header row, describes a particular entity of the concept associated with this table.

The methods (Douglas et al., 1995; Embley et al., 2005; Tijerino et al., 2005; Wang et al., 2012) use mainly domain knowledge about natural language content of tables. However, it is not always sufficient in practice. There are many cases when the table understanding additionally requires an analysis of spatial and graphical information from tables.

An opposite domain-independent method to extract information from HTML tables is offered by Gatterbauer et al. (2007). It is based on the analysis of only spatial and style information in the CSS2 (Cascading Style Sheets Level 2) format. In particular, they propose to carry out the interpretation of the tables (recovery of semantic relationships) based on assumptions about style information designed for a set of the most common types of web-tables.

Pivk (2006) and Pivk et al. (2007) present a methodology and TARTAR system for automatic transforming HTML tables of three typical types into logical structured form (semantic frames) that is intended for using with an inference engine for the query answering and ontology generation. The methodology and system are also independent of domain knowledge. They are based on heuristics on layout and text content of a table.

The paper (Kim & Lee, 2008) proposes a method for extracting logical structures (where semantic relationships between attributes and values are presented as tree) from HTML tables and transforming them into a XML representation. Their method is restricted by five types of tables. Kim and Lee (2008) use an analysis of spatial, style and natural language information from a table based on embedded rules and regular expressions.

A detailed description features of several others methods, in particular, (Chen, Tsai, & Tsai, 2000; Hu, Kashi, Lopresti, & Wilfong, 2000; Hurst, 2000; Pinto, McCallum, Wei, & Croft, 2003; Yoshida, Torisawa, & Tsujii, 2001), for functional analysis, structural analysis, and interpretation of a table is given in the paper (e Silva et al., 2006). As a rule, they are based on using some assumptions about table structures in steps of functional or structural analysis of a table. Those assumptions limit a class of tables which can be understood by these methods with a high precision and recall.

## 3. Class of processed tables

Now, the large volume of unstructured tabular information is presented in high-level document formats, such as Excel, Word, and HTML. The possibilities and constraints of the table presentations in these formats are similar. They allow to present the following information about a table:

- Positions of a cell in row and column coordinates;
- Merged cells (e.g. attributes COLSPAN and ROWSPAN in HTML);
- Cell style (border style, content placement, text metrics, etc.);
- Content of a cell (text, images, etc.).

However, each of the formats has its own features. So a cell can contain other tables in Word and HTML. But Excel does not supported it. HTML allows using the attributes HEADERS, SCOPE of the tags TD and TH to define relationships between headers and values. Excel determines one of the primitive data types (NUMERIC, DATE, STRING, etc.) for cell content.

[1] WordNet, http://wordnet.princeton.edu.

[2] Probase, http://research.microsoft.com/en-us/projects/probase.