Towards Lightweight Representation of the Table Semantics for the Cross-Context Information Exchange

Alexey Shigarov¹, Vasiliy Khristyuk¹, Evgeniy Cherkashin¹, and Shuo Yang²

¹ Matrosov Institute for System Dynamics and Control Theory of SB RAS, 134 Lermontov st., Irkutsk, Russia, 664033 shigarov@icc.ru,

Shigaroverce.ru,

WWW home page: http://td.icc.ru ² School of Computer Science and Cyber Engineering, Guangzhou University, 230 Wai Huan Xi Road, Guangzhou, China, 510006 yangshuo@gzhu.edu.cn

Abstract. This addresses representation of the table semantics for the cross-context information exchange. The tables we consider have an arbitrary cells structure represented in a machine-readable format. For example, tables can be contained in electronic documents, such as a spreadsheet or a web-page, and, typically, they are not accompanied by semantics enabling their automatic interpretation. In spite of the existing variety of formalisms for representation of the table semantics, most of them they are fairly inefficient in terms of the user efforts required for the semantic annotation. We outline a new approach to the lightweight representation of the table semantics. We stay on the interpretation level that provides the inference of the semantics of atomic data items of a table from a description of data groups expressed by the syntax of the table. We expect that, in the future, implementation of our approach can reduce the complexity and volume of the table semantics required for the cross-context information exchange, as well as the user efforts aimed at annotating tabular data.

Keywords: Table Semantics · Table understanding · Table interpretation · Semantic interoperability · Information exchange · Spreadsheet.

1 Introduction

Nowadays, the volume of electronic documents participating in the information exchange and transmission continues to significantly increase. The interpretation of a document depends on its context (historic, national, domain, organizational, etc.). For example, if a table in a financial report is titled with

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

"FY2020", then, in Russia, this might mean some data of a fiscal year starting from January 1, 2020; however, in the United States, we should infer a different year period starting on October 1, 2019. This complicates the document processing and understanding within the information exchange.

The cross-context exchange of electronic documents implies that a document is represented in one context and transmitted for processing in another context [13, 16]. The ability of computer programs to exchange information with unambiguous and shared meaning is referred to as the *semantic interoperability* [2]. Particularly, the semantic interoperability should be provided when a source context differs from a target one. The use of semantic markup makes documents readable and interpretable not only by humans but also by computer programs. Such markup plays a key role in enabling semantic interoperability for the information exchange.

In recent years, many studies were devoted to some issues of semantic interoperability (e. g. [6, 12, 14, 15]. However, the problem of the cross-context exchange of tabular documents remains open in general. The arbitrary tables with an explicit syntactic structure of cells are one of the main ways to present data in electronic documents. For example, they can be contained in a spreadsheet, a rich text document, or a hypertext of a web-page. Typically, such tables are not accompanied by explicit semantics needed for their automatic interpretation. The interpretation of such documents is complicated due to the variety of forms for representing the table syntax and semantics. The actual challenges are the extraction of the semantic components from documents, the context neutralization, and the unambiguous interpretation of transmitted data, as well as the representation of the semantic structure of tables.

We outline a new 3-Level Table Object Model (3L-TOM) for the lightweight description of tables on the interpretation level. It is assumed that the model will enable the inference of atomic data items from categories, i. e. groups of data items. We expect that our approach can reduce the complexity and volume of the semantics required for the annotation and interpretation of arbitrary tables in the cross-context exchange of electronic documents.

2 Problem Statement

The representation of the table semantics for the cross-context exchange of electronic documents is a complicated problem. Such representation should primarily enable both the semantic annotation of arbitrary tables and the interpretation of tabular data in a target context. Additionally, it has to reduce the complexity and volume of the table semantics transmitted in the information exchange.

We begin with a discussion about the concept of an arbitrary table. Hinterberger [3] defines a table both as a data structure to organize the tuples of relation and as an arrangement of data in rows and columns. On the one hand, a table can represent relational data in a grid of cells (e. g. contingency tables used in statistics). On the other hand, a table can be used as a way of visual communication to arrange data items, even when there are no relationships between them (e. g. the grid layout used in web pages).

We refer to an *arbitrary table* presented in a document as a way of visual communication for arranging interrelated data items in a grid of cells. The data items represented in an arbitrary table are divided into two functional types: (i) entries and (ii) labels. The *entries* are values of data, while the *labels* are considered as keys or attributes for addressing the values of data conceptually. In comparison with a relational table, where its tuples contain only values but its schema distinguishes attributes, an arbitrary table represents both values and attributes in one syntax structure, the grid of cells. A relational table typically describes instances of one conceptual entity, while an arbitrary table often includes labels of several conceptual entities.

The arbitrary tables are mainly intended to be understood by humans. Typically, they lack explicit semantics needed for computer programs to interpret them as intended by their authors or as required by an application. For example, a table presented in a spreadsheet can be read and understood by humans as relational information, but such a table is just a grid of cells hidden in another grid of cells on a sheet for computers. The syntax of the arbitrary tables allows expression of semantics implicitly via various functional arrangements of data items in cells, as well as the formatting variety of both cells and text. In general, the semantic annotation of an arbitrary table requires execution of the main steps of the table understanding.

Now, consider two approaches to representation of table semantics that we refer to as heavy and lightweight representation, respectively. The first approach describes semantics heavily on the level of atomic data items. Alternatively, the second one is a lightweight description of data groups, such as categories. The volume of the heavy representation grows linearly with respect to the amount of atomic data items, while the volume of the lightweight representation depends only on a number of groups, such as data items. Even the complex multidimensional tables rarely contain more than a dozen categories, which can be interpreted as separate groups. For example, suppose that we need to describe the semantics of a table containing 10,000 data items, of which 1,000 are labels of one category (L_1) , 1,000 are labels of another category (L_2) , and 8,000 are entries of the third category (E). Moreover, each entry of E is addressed by two labels: one of L_1 , the other of L_2 . In this case, a heavy representation should describe 10,000 objects and 16,000 "entry-label" relationships between them. On the other hand, such a table contains only 3 groups $(L_1, L_2, \text{ and } E)$. Therefore, lightweight semantics is limited to describing only 3 objects and 2 relations between them.

The existing formalisms for representing tabular data such as Wang's model [10, 11], Hurst's model [4], 2-Level Table Object Model (2L-TOM) [8, 9], and Relational Data Model (RDM) [1,5] are fairly inefficient for the semantic annotation of arbitrary tables in terms of the user efforts. The first three models [8,9] might be used to describe the semantics of arbitrary tables only in a heavy manner. This case requires mapping of an arbitrary table to its atomic data items. The RDM [1,5] allows specification of a table via a conceptual schema. A volume of such representation depends on the number of table columns, so it can be considered as a lightweight description. However, this model strongly restricts a syntactic structure by the definition of relational tables [5], when each column corresponds to a labeled attribute and each row is a tuple of values. Note that many tables presented in electronic documents have a more complex syntactic structure. Therefore, RDM does not apply to the semantic annotation of arbitrary tables in the general case.

We believe that it is possible to develop a model for the lightweight representation of the table semantics on the interpretation level, which can be utilized for a substantial range of arbitrary tables. In comparison with the existing formalisms, the model we aim to build should enable the inference of atomic data items from their categories.

3 Solution Outline

We outline a new 3-Level Table Object Model (3L-TOM) for representing the structure of an arbitrary table on the following three levels: (i) syntactic, (ii) semantic, and (iii) interpretation. The 3L-TOM model extends the 2L-TOM by adding the third level, a lightweight representation of categories for the table interpretation. The 2L-TOM was introduced in [8,9] and implemented as a software [7] to provide the syntax and semantics of arbitrary tables.

The first level of the 2L-TOM describes syntactic objects of a table, such as a layout, formatting, and text of cells. This level should comply with the capabilities and limitations of contemporary table formats such as Excel and HTML. The second level defines data items of two functional types: (i) entries and (ii) labels. All semantic objects are separated into two or more groups. At least two groups are corresponding to the different functional types. Each entry can be associated with one label of each group. Labels of the same group can be associated with each other by parent-child relationships. Both entries and labels are typically read and converted from the text of some cells. To simplify the presentation of the model, we use the following assumptions: the properties of syntactic objects (layout, formatting, and text) are attributed to the semantic objects produced from the corresponding cells. For example, when a label was created as a result of text reading from a cell at the address A1:B2 of a source spreadsheet, then we say that this label is located at the address A1:B2. Note that the semantic object does not have syntactic properties directly, but they can be inferred from the associated syntactic objects. The semantic level can be made context-independent by neutralizing the context of values read from the syntactic level.

The third level of the 3L-TOM interprets a table by specifying semantic groups of data items. A semantic group is a set of data items which belong to the same functional type and the same category of an external vocabulary (e. g. DBpedia³, YAGO⁴, and Wikidata⁵). Each semantic group determines a set of operations for generating its data items from syntactic objects of a table and context. Each pair of groups, where one is a set of entries and another is a set of labels, determines a set of operations for coupling their data items by the entry-label relationships. A group of labels can also determine a set of operations for coupling its labels by the label-label relationships. The interpretation level provides the automatic inference of the semantics from the syntax of a table. It can also serve to validate both the semantics and the syntax of a table.

While the syntactic and semantic levels can be implemented by 2L-TOM, the interpretation level is supposed to be based on describing semantic groups of data items not provided by this model. To represent the interpretation level of 3L-TOM, the three-level table model, we propose a design of a novel language for a lightweight description of the table semantics, hereinafter TSDL (Table Semantic Description Language). This language aims at reducing the volume and the complexity of semantic annotation of tables through a lightweight description of entire groups, instead of a heavy description of atomic data items.

The design of TSDL is based on predefined operations for the context neutralization of tabular data. They serve to cleanse data items read from table cells, as well as to free them from the context-dependency. A data item is read from the textual content of one or more cells. Its value can be modified through various transformations such as string processing, type conversion, aggregation, etc. Some transformation pipelines can be composed of several operations. It is assumed that one or more of such pipelines provides generation of all data items of a group context-independently.

Another kind of the predefined operations that the design of TSDL is intended to include is linking data items with each other by inner relationships of the table structure. These operations correspond to some general features of the table layout recommended by typographical standards and observed in many documents. Our approach divides these operations into two types per two kinds of inner relationships: entry-label and label-label. The first type of operations uses methods for the inference of entry-label relationships from a pair of groups where one contains entries while another consists of labels. The second type determines methods for the inference of parent-child relationships between the labels of the same group.

The operations for the entry-label linking are based on the following general features of the table layout:

- BY_ROW / BY_COLUMN / BY_CELL, an entry is associated with a label when they are originated from cells placed in the same row, column, or cell, respectively.
- BY_INDEX, an entry is associated with a label when they are originated from cells read in the same order.

³ https://dbpedia.org

⁴ https://yago-knowledge.org

⁵ https://www.wikidata.org

- BY_ADDRESS, an entry is associated with a label originated from a cell at a specified address.
- BY_SINGLE, all entries are associated with a single label of a semantic group.

Note that these operations can be parametrized to specify some control information, for example, a direction for seeking a cell of the label relative to a cell of the entry, a shift in the reading order of cells, a cell address.

The operations for the label-label linking engage the following general features of the table layout:

- BY_NESTING, child-parent labels are originated from cells located in adjacent rows (columns), the child cell is nested in the parent cell by columns (rows).
- BY_INDENTATION, child-parent labels are originated from cells located in the same column, the text of the child label is indented relative to the text of the parent cell.
- BY_EMPHASIZING, child-parent labels are originated from cells located in the same column, the text of the labels is highlighted by the different font formatting.
- BY_ALIGNMENT, child-parent labels are originated from cells located in the same column, the text of the labels is highlighted by the different alignment.

We expect that some additional features of the table layout can be identified over time. Therefore, the lightweight description should be extensible by new operations for linking groups. The language might be designed as functional. In this case, the operations of the context neutralization and the linking of semantic groups can be expressed as function calls.

4 Conclusion

In the future, we expect that implementation of the approach proposed might reduce the complexity and volume of the table semantics required for the crosscontext information exchange, as well as the user efforts aimed at annotating tabular data.

The 3L-TOM model could enable implementation of the visual annotation of a table on the level of data groups described in both a syntactic and semantic side. Typically, all data items of one group are originated from cells located in one functional region (adjacent rows and/or columns). For example, pivot tables often place all labels of one category into either one row of a head or one column of a stub, while all entries are placed into body cells. This layout feature can be used to visually annotate semantic groups by selecting the corresponding functional regions of cells. We expect that such annotation allows creation of the semantic markup with minimal efforts of end-users.

Another consequence arising from the implementation of 3L-TOM is the possibility to recover semantics, the 2nd level of an instance of 3L-TOM, from

syntax, the 1st level of the instance of 3L-TOM. Interpretation of the TSDL descriptions enables either the inference of atomic data items of a table when they are absent or their validation when they are present in the instance of 3L-TOM. This may significantly reduce the volume of tabular data transmitted from one context to another. One lightweight description of semantics can be applied to a set of tables with the same layout but with different content. Potentially, the validation prevents unacceptable modifications of the table layout and content in data collection tasks, as well as protection of tabular data against potential damage in document exchange tasks.

Moreover, atomic data items restored by using the TSDL descriptions can be represented as linked data. Such a format complies with the Linked Data⁶ principles that allow creation of semantic objects published in the form of hydpertext, at the same time linking them to elements of other documents and objects. It is possible to construct some rules for mapping table semantics to the RDF⁷ graph with the concretization of syntax via the standards like RDFa/XML⁸ and Turtle⁹. The popular common-sense knowledge graphs (e. g. DBpedia, YAGO, and Wikidata), which can be used to describe table semantics, support RDF as a standard knowledge representation format. This allows integration of the linked data generated from tables with open external vocabularies in a common format. The proposed translation of table semantics into RDF will simplify the utilization of such data in some target applications, since RDF is supported by the majority of the ontological modeling tools and, de facto, it is the most used tool for representing linked data and ontologies.

Summarizing the above, our approach can be implemented in the future by the development of the following tools: (i) a model representing tables on three levels (syntactic, semantic, and interpretation), (ii) a formal language for a lightweight description of the table semantics, (iii) a visual annotator for document tables to get a lightweight description, (iv) a validator of the table syntax and semantics, (v) a generator of linked data from tables by their lightweight description. We believe that these results will be useful in applications of cross-context exchange of tabular documents in various fields (egovernment, e-healthcare, e-commerce, etc.).

5 Acknowledgment

This work was supported by the Basic Research Program of the Siberian Branch of the Russian Academy of Sciences, Project IV.38.1.2, Registration No. AAAA-A17-117032210079-1.

⁶ https://www.w3.org/wiki/LinkedData

⁷ https://www.w3.org/RDF

⁸ https://www.w3.org/TR/rdfa-core

⁹ https://www.w3.org/TR/turtle

References

- Embley, D.W.: Relational model. In: Encyclopedia of Database Systems, pp. 3149– 3154 (2018). https://doi.org/10.1007/978-1-4614-8265-9_306
- Heiler, S.: Semantic interoperability. ACM Comput. Surv. 27(2), 271—-273 (1995). https://doi.org/10.1145/210376.210392
- Hinterberger, H.: Table. In: Encyclopedia of Database Systems, pp. 3873–3874 (2018). https://doi.org/10.1007/978-1-4614-8265-9_1373
- Hurst, M.: Towards a theory of tables. Int. J. Doc. Anal. Recog. 8(2-3), 123–131 (2006). https://doi.org/10.1007/s10032-006-0016-y
- Johnston, T.: Chapter 3 the relational paradigm: mathematics. In: Bitemporal Data, pp. 35–41 (2014). https://doi.org/10.1016/B978-0-12-408067-6.00003-6
- Qin, P., Guo, J.: A novel machine natural language mediation for semantic document exchange in smart city. Future Generation Computer Systems 102, 810–826 (2020). https://doi.org/10.1016/j.future.2019.07.028
- Shigarov, A., Khristyuk, V., Mikhailov, A.: Tabbyxl: Software platform for rulebased spreadsheet data extraction and transformation. SoftwareX 10, 100270 (2019). https://doi.org/10.1016/j.softx.2019.100270
- Shigarov, A., Khristyuk, V., Mikhailov, A., Paramonov, V.: Tabbyxl: Rule-based spreadsheet data extraction and transformation. In: Information and Software Technologies. vol. 1078 CCIS, pp. 59–75 (2019). https://doi.org/10.1007/978-3-030-30275-7_6
- Shigarov, A.O., Mikhailov, A.A.: Rule-based spreadsheet data transformation from arbitrary to relational tables. Information Systems 71, 123–136 (2017). https://doi.org/10.1016/j.is.2017.08.004
- 10. Wang, X.: Tabular abstraction, editing, and formatting. Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada (1996)
- Wang, X., Wood, D.: A conceptual model for tables. In: Principles of Digital Document Processing. vol. 1481 LNCS, pp. 10–23 (1998). https://doi.org/10.1007/3-540-49654-8_2
- Yang, S., Wei, R.: Tabdoc approach: An information fusion method to implement semantic interoperability between IoT devices and users. IEEE Internet of Things Journal 6(2), 1972–1986 (2019). https://doi.org/10.1109/JIOT.2018.2871274
- Yang, S., Wei, R.: Semantic interoperability through a novel cross-context tabular document representation approach for smart cities. IEEE Access 8, 70676–70692 (2020). https://doi.org/10.1109/ACCESS.2020.2986485
- 14. Yang, S., Wei, R., Guo, J., Xu, L.: Semantic inference on clinical documents: combining machine learning algorithms with an inference engine for effective clinical diagnosis and treatment. IEEE Access 5, 3529–3546 (2017). https://doi.org/10.1109/ACCESS.2017.2672975
- Yang, S., Guo, J., Wei, R.: Semantic interoperability with heterogeneous information systems on the internet through automatic tabular document exchange. Information Systems 69, 195–217 (2017). https://doi.org/10.1016/j.is.2016.10.010
- 16. Yang, S., Wei, R., Shigarov, A.: Semantic interoperability for electronic business through a novel cross-context semantic document exchange approach. In: Proc. ACM S. on Doc. Eng. pp. 28:1–28:10 (2018). https://doi.org/10.1145/3209280.3209523