Semantic Integration of Heterogeneous and Complex Spreadsheet Tables*

Sara Bonfitto[1][0000−0002−9883−5561]

Computer Science Dep., Università di Milano, Via Celoria 18, Milano
sara.bonfitto@unimi.it

Abstract. A great number of companies and institutions use spreadsheets for managing, publishing and sharing their data. Though effective, spreadsheets are mainly designed for being interpreted by humans, and the automatic extraction of their content and interpretation is a complex task. The task becomes even harder when tables present different kinds of mistakes and their layout is complex. In this paper, we outline the approach that we wish to develop during the PhD for answering the research question "how to semi-automatically extract coherent semantic information from heterogeneous and complex spreadsheets?".

Keywords: Heterogeneous Spreadsheet Tables · Semantic Table Interpretation · User Interfaces · Machine Learning.

1 Introduction

Recently, our research group was involved in the problem of integration of heterogeneous spreadsheet files that a debt collection agency daily receives from local authorities (e.g. municipalities, tax agency) containing batches of thousand of invoices to be rescued. These spreadsheets are big, heterogeneous and do not follow any standard format or notation (Fig. 1 shows an example). The first row reports the column headers, however, the access keys are not always present and do not follow any specific format. Data occurring in the same column sometimes adhere to different types. For example, the column SSN/VAT contains different data types (actually expressing that invoices can be titled to individual citizens or to companies). Sometimes columns present strings from which different kinds of information can be extracted, as in the case of the "address" column, where different alternative patterns represent the street name, and street/apartment number. Blanks and semi-blank rows can occur in the main table. Semi-blank rows usually contain totals or aggregated data. Blank rows are sometimes used for aesthetic reasons while others for separating rows representing correlated invoices. Indeed, the information about an invoice is not always contained in a single table row. For example, in Fig. 1 there is a correlation among two rows (the fourth row contains the reference to the legal representative associated with

* PhD Advisor Prof. Marco Mesiti

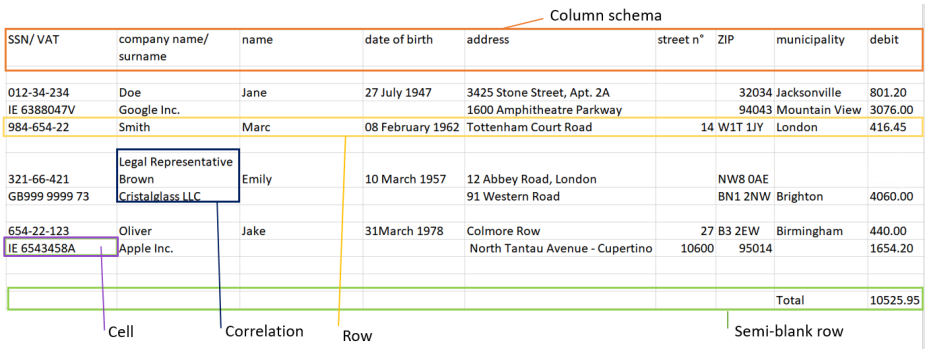

Fig. 1. A spreadsheet example

the invoice in the fifth row). This kind of correlation can be expressed by following different patterns. Last but not least, the information contained in these spreadsheets can contain different kinds of typographical, grammatical and miscalculation errors. The variability of organizations of these spreadsheets prevent the use of well studied approaches for table understanding (e.g. [3]), data repairs and extraction (e.g. [5]), data transformation (e.g. [6]), programming by example (e.g. [4]), and semantic characterization of the information (e.g. [8]). Standard approaches for NLP cannot be applied on short texts like the one that can occur in spreadsheets for extracting patterns. We believe that a completely automatic approach that exploits sophisticated machine learning (ML) techniques cannot properly be used in this context. A semi-automatic approach can be devised to support the user during the process of data cleaning, transformation and semantic characterization that involve the user in the loop in order to tune the prediction system depending on the feedback obtained while processing new spreadsheets. Users need to be supported by easy-to-use graphical interfaces for correcting mistakes and improve the overall performance of the system.

In order to reach this goal, we propose the adoption of a three-phase approach. Phase I is responsible for the spreadsheet cleaning, the identification of the column types and the synthetic error correction. Phase II aims to create a semantic characterization of the table content to be extracted from the spreadsheet and relies on the use of a domain Ontology and, when possible, a Knowledge Base. Phase III relies on the identification of semantic constraints and assertions that need to be checked and maintained on the considered Ontology. The purpose of this phase is to point out semantic mistakes that can be fixed on the RDF representation of spreadsheet tables.

2 The three-phase approach

Phase I: Table Identification and Cleaning. The main purpose of this phase is to correct syntax errors occurring in the data, identifying the correlations existing among table rows, and identifying basic types of each column.

For identifying correlation among table rows, we wish to adopt a declarative pattern-based language for specifying when a correlation exists. Moreover, we wish to develop interfaces for further supporting the users in their manual identification and thus learning new patterns for the interaction. Moreover, we wish to develop a multi-label classification approach for the identification of the cell and column types. Several basic types, domain-specific types and also pattern-based types will be supported. Patterns will be exploited for extracting values from complex strings like for example the address "12 Abbey Road, London" in Fig. 1. A multi-label approach is considered for facing situations like the column "company name/surname" that contains both the company name or the citizen surname. The automatically identified types, however, can contain errors due to the occurrence of mistakes in the data. Therefore specific interfaces should be developed for their easy correction. Moreover, the large amount of invoices to be processed requires the adoption of solutions that apply a single correction to many invoices at the same time.

Phase II: Semantic Characterization of table Content. The aim of Phase II is to provide a *semantic meta-model* description of the spreadsheet tables by means of annotations w.r.t. a Domain Ontology. Even if many approaches have been proposed so far for this problem, in our research we wish to face the variability of data types identified in the first phase that usually is not considered. Moreover, the semantic meta-model should be coupled with a graphical representation that makes easier to the user checking the automatically generated model and correcting mistakes when needed. Moreover, a ML algorithm will be applied for learning annotations relying on previously established mappings. The user can also change manually the annotations, these modifications should be exploited for tuning the predictions. Our semantic meta-model is inspired by the one used in Karma [7] but differs from it because it is created starting from the types identified in the first phase and allows the extraction of several data from a single column (while Karma only makes a 1:1 correspondence from the data in the spreadsheet to the Ontological concepts).

Phase III: Verification of Semantic Constraints and Assertions. The semantic meta-model is finally used for the automatic extraction and transformation of data in an RDF format according to the domain Ontology. In this phase, we wish to use the semantic constraints and assertions identified on the Ontology to point out semantic mistakes. For example, the total debt amount in Fig. 1 is correct and corresponds to the sum of the single debt imports, the zip code of an address can be validated against the municipality. These are semantic constraints w.r.t. the syntactic constraints identified in Phase I.

3 Concluding Remarks

In this paper, we outlined our main research question and the related problems that should be faced in the next two years of the PhD program.

At current stage, we have started working on a survey on related works in the context of table understanding and semantic interpretation of tables [2]. In this survey, we have outlined the different phases in which the table understanding problem can be organized (localization, segmentation, functional and structural analysis and integration) and presented the main approaches proposed in the last fifteen years. Moreover, an initial solution for the first phase is proposed in [1] by introducing a methodology for determining the value/column types contained in CSV tables that exploits a multi-label prediction algorithm that has been trained on a simulation of typical data available in the considered domain that takes into account the errors occurring in data. This automatic approach has been combined with graphical user interfaces with which the user can check the predicted types and modify them when needed. The modifications can be applied at type-level, thus many values can be modified by a single specification. This initial activity needs to be further enhanced for identifying correlated rows and cells containing an aggregation of other cells and also functional relationships that need to be preserved on data (e.g. the occurrence of an SSN requires the presence of name and surname of an individual).

We are currently working on the second phase of the approach by specifying the semantic-description of a spreadsheet table and its graphical representation. Moreover, an approach for the semi-automatic construction of the model is evolving that takes into account the previously specified meta-models and similarity measures for evaluating their adequateness to the new scenario.

References

1. Bonfitto, S., Cappelletti, L., Trovato, F., Valentini, G., Mesiti, M.: Semi-automatic column type inference for csv table understanding. In: 47th Int. Conf. on Current Trends in Theory and Practice of Computer Science. pp. 535–549. Springer (2021). https://doi.org/10.1007/978-3-030-67731-2
2. Bonfitto, S., Casiraghi, E., Mesiti, M.: Table understanding approaches for extracting knowledge from heterogeneous tables. WIREs Data Mining and Knowledge Discovery (2020), to appear
3. Holeček, M., Hoskovec, A., Baudiš, P., Klinger, P.: Table understanding in structured documents. In: Proc. of Int'l Conf. on Document Analysis and Recognition Workshops (ICDARW). vol. 5, pp. 158–164 (2019)
4. Jin, Z., Anderson, M.R., Cafarella, M., Jagadish, H.V.: Foofah: Transforming data by example. In: Proc. of ACM SIGMOD. p. 683–698 (2017). https://doi.org/10.1145/3035918.3064034
5. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive visual specification of data transformation scripts. In: ACM Human Factors in Computing Systems (CHI). p. 3363–3372 (2011). https://doi.org/10.1145/1978942.1979444
6. Shigarov, A., Khristyuk, V., Mikhailov, A., Paramonov, V.: Tabbyxl: Rule-based spreadsheet data extraction and transformation. SoftwareX pp. 59–75 (10 2019). https://doi.org/10.1007/978-3-030-30275-7_6
7. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: Learning the semantics of structured data sources. Journal of Web Semantics **37**, 152–169 (2016)
8. Zhang, Z.: Effective and efficient semantic table interpretation using tableminer$^+$. Semantic Web **8**(6), 921–957 (2017). https://doi.org/10.3233/SW-160242