Semantic Analysis of Tabular Data

Subjects: Computer Science, Information Systems

Contributor: Madina Mansurova , Vladimir Barakhnin , Assel Ospan , Roman Titkov

Understanding the semantics of a table makes it possible to find relationships between various data, form a general picture of the contents of the table, and extract information in a convenient form for further processing.

semantic analysis table interpretation

1. Introduction

Currently, data collection and analysis are an integral part of the vast majority of fields of activity. Information can be presented in various formats, such as text, tables, and images. At the same time, one of the critical tasks is the analysis of tables that contain large amounts of information.

To extract the information needed from a table, one needs to analyze it and understand the semantics, including the content type of each cell and the relationships between them. It is important to consider that the structures of tables can differ significantly from each other.

2. Semantic Analysis of Tabular Data

Nowadays, there are several solutions that allow for the semantic analysis of tables.

TableMiner+ [1] is a tool designed to automatically extract table structures and content from HTML documents. It detects tables, extracts headers, cells, columns, and rows, and determines the data types in tables. The tool uses an iterative approach to refine the quality of data extraction. However, it may run into problems when working with complex tables, such as tables with graphics or formulas. TableMiner+ works based on certain table formatting rules and requires adjustments for tables that do not follow those rules. It only handles PDF and HTML formats. Despite its limitations, it is effective for semantic data extraction from tables and has applications in areas such as medical statistics and data science.

QTLTableMiner++ [2] extracts semantic information from tables in articles related to genetics. It uses natural language processing and machine learning to extract information about genetic variants from tables. The tool defines tables of genetic variant data and links them to the corresponding genetic loci. It is based on the Apache Solr search engine and can index and search text data in multiple languages. However, its use is limited to certain table formats from a particular resource. Overall, it is a powerful tool for extracting genetic data, but it has limitations, such as a limited table format and a domain-specific codebase.

Qurma is a pipeline for extracting tables, interpreting and enriching knowledge graphs from various sources, based on the Camelot library [3]. Qurma efficiently processes HTML pages, PDFs, and images initiated by a user-provided document URL to retrieve a table. The result is an optimized dataset that can be exported as CSV, JSON, or attribute values.

Qurma's conceptual architecture is based on the Clean Architecture paradigm, with a focus on controlled interfaces and a layered structure.

MantisTable is a public tool for semantic annotation of tables from scientific publications [4]. It identifies objects in table cells and links them to knowledge bases like DBpedia using string matching, word embedding, and machine learning. The tool was evaluated on 100 tables from scientific articles, comparing its abstracts with those of experts. It demonstrated high F1 values for object recognition and linking. Compared to other tools, MantisTable showed the best F1 values for both tasks. Its intuitive interface is highly acclaimed by users. Semantic annotations from MantisTable can improve the search and retrieval of scientific content.

Berners-Lee et al. introduced the concept of the Semantic Web, where data is given a well-defined meaning that allows computers and people to work together [5]. OWL is a key component of this vision, providing a rich language for defining structured web ontologies. The disadvantages of this method are (1) difficult to implement, as the vision of the Semantic Web requires significant changes to the existing web infrastructure, and (2) dependence on widespread adoption; for the Semantic Web to be effective, most web content creators must adopt standards, (3) the possibility semantic inconsistencies; different creators may interpret values differently, resulting in semantic inconsistencies.

Hurst, M. discussed the problems and techniques involved in interpreting tables in documents, emphasizing the need to understand both structure and semantics [6]. It has the following disadvantages: (1) limited to document-based tables—may not be efficient for web tables or dynamic tables; (2) problems handling different table structures and formats.

In [7], the authors proposed an approach to interpreting tables using linked data and ontologies. Their method focuses on creating RDF triplets from tables. However, this method has a dependency on the quality and completeness of ontologies and potential problems when matching tables with existing ontologies, especially if they are not aligned.

The authors of [8] developed integrated machine learning with semantic technologies to interpret and annotate tables where there are difficulties with dependence on training data since the quality and quantity of training data can affect performance.

A tool called Table2OWL has been developed to transform tables into OWL ontologies. It provides a semi-automatic way to convert tabular data to semantic web formats [9], where manual intervention is required, which can be time-consuming, and errors are possible when converting tables in OWL ontologies. Elaborating further, the complexity arises due to diverse data structures in tables, requiring precise class and property decisions in OWL. Often, tables possess inconsistencies or ambiguous data, which, when mapped to ontologies, can lead to representation errors. Moreover, preserving the semantic integrity of information during conversion demands expert judgment, especially when interpreting relational nuances from tables. Post-conversion, thorough validation against the source table is indispensable. Despite Table2OWL's assistance, manual oversight, driven by domain expertise, remains crucial to ensure accuracy.

Limaye et al. introduced a system of semantic table annotation, focusing on linking table cells to DBpedia objects [10]. The disadvantage of this method is the limitation of binding to DBpedia objects: it may not cover all possible semantic annotation, and problems with handling ambiguous or obscure table cell contents.

Poggi et al. discussed data access based on OBDA ontologies, where relational data sources are practically integrated with the ontology, providing a semantic representation of data [11]. Here, the authors face the difficulty of integrating relational data sources with ontologies.

Another study [12] presented WebTables, a system that retrieves and indexes tables from the Internet, providing a semantic search of a huge amount of tabular data. However, there are also problems with retrieving and indexing various web tables, and there may be semantic inconsistencies during searches.

Bhagavatula et al. proposed methods for linking tables to the correct data models, which allows better understanding and querying of tabular data [13]. The work has been successfully applied to tables but faces the following difficulties: (1) dependency on existing data models; if the models are not exhaustive, interpretation can be limited; (2) problems in associating different tables with the correct models.

To address the problems of interpreting tables, Venetis et al. discussed the problems of interpreting tables, especially when working with web tables, which can be noisy and heterogeneous [14]. Although this work highlights the problems, it does not offer comprehensive solutions.

References

1. Zhang, Z. Effective and Efficient Semantic Table Interpretation using TableMiner+. Semant. Web 2016, 8, 921–957.
2. Singh, G.; Kuzniar, A.; van Mulligen, E.; Gavai, A.; Bachem, C.; Visser, R.; Finkers, R. QTLTableMiner++: Semantic mining of QTL tables in scientific articles. BMC Bioinform. 2018, 19, 183.
3. Nugumanova, A.; Apayev, K.; Baiburin, Y.; Mansurova, M.; Ospan, A. Qurma: A table extraction pipeline for knowledge base population. J. Math. Mech. Comput. Sci. 2022, 114.
4. Lamy, J.B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif. Intell. Med. 2017, 80, 11–28.
5. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. Sci. Am. 2001, 284, 29–37.
6. Hurst, M. The Interpretation of Tables in Texts. Ph.D. Thesis, University of Edinburgh, Edinburgh, UK, 2000.
7. Mulwad, V.; Finin, T.; Joshi, A. T2LD: Interpreting and Representing Tables as Linked Data. In Proceedings of the Poster and Demonstration Session at the 9th International Semantic Web Conference, Shanghai, China, 9 November 2010; Available online: https://www.researchgate.net/publication/221466623_T2LD_Interpreting_and_Representing_Tables_as_Linked_Da (accessed on 9 November 2010).
8. Syed, Z.; Finin, T.; Mulwad, V. Exploiting a Web of Semantic Data for Interpreting Tables. In Proceedings of the Second Web Science Conference, Raleigh, NC, USA, 26–27 April 2020; Available online: https://www.researchgate.net/publication/228806445_Exploiting_a_Web_of_Semantic_Data_for_Interpreting_Table (accessed on 27 April 2010).
9. Jannach, D.; Shchekotykhin, K.; Friedrich, G. Automated ontology instantiation from tabular web sources—The AllRight system. J. Web Semant. 2009, 7, 136–153.
10. Limaye, G.; Sarawagi, S.; Chakrabarti, S. Annotating and Searching Web Tables Using Entities, Types and Relationships. Proc. VLDB Endow. 2010, 3, 1338–1347.

11. Poggi, A.; Lembo, D.; Calvanese, D.; De Giacomo, G.; Lenzerini, M.; Rosati, R. Linking Data to Ontologies. In Journal on Data Semantics X; Spaccapietra, S., Ed.; Springer: Berlin/Heidelbrg, Germany, 2008; Volume 4900.
12. Cafarella, M.J.; Halevy, A.; Wang, D.Z.; Wu, E.; Zhang, Y. WebTables: Exploring the Power of Tables on the Web. Proc. VLDB Endow. 2008, 1, 538–549.
13. Bhagavatula, C.S.; Noraset, T.; Downey, D. Methods for Exploring and Mining Tables on Wikipedia. In Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, Chicago, IL, USA, 11 August 2013.
14. Venetis, P.; Halevy, A.; Madhavan, J.; Pasca, M.; Shen, W.; Wu, F.; Miao, G.; Wu, C. Recovering Semantics of Tables on the Web. Proc. VLDB Endow. 2011, 4, 528–538.

Retrieved from https://encyclopedia.pub/entry/history/show/116233