УДК 004.67
Гашинський Р. – ст. гр. СН-41, Мельник Н. – ст. гр. СІм-52
Тернопільський національний технічний університет імені Івана Пулюя

АВТОМАТИЗОВАНИЙ АНАЛІЗ СХЕМИ ДОКУМЕНТІВ З ЧАСТКОВО-СТРУКТУРОВАНИМИ ДАНИМИ

Науковий керівник: асистент Кравчук Г. Б.

Gashynskyi R., Melnyk N.
Ternopil Ivan Puluj National Technical University

AUTOMATED ANALYSIS OF THE SCHEMA OF DOCUMENTS WITH SEMI-STRUCTURED DATA

Supervisor: assistant Kravchuk H.

Ключові слова: дані, метадані, менеджмент даних
Keywords: data, metadata, data management

In today's information society, where data rapidly increases in volume and becomes more diverse, ensuring high quality of the data is important to provide solid foundations for management and business decisions. Most organizations work with data that can be both a business object and a result of the activities of the organizations themselves. If an organization can't trust data to meet business needs, all efforts spent on collecting, storing, protecting, and making it available will be wasted [1].

Usually data comes in the form of documents from different sources and in different formats. The lack of a common document schema can create significant problems for data processing and further analysis. The problem of semi-structured data [2] manifests itself in differences in data storage formats, inconsistencies in the nomenclature of columns in tables. For example, one data source may use the date format "mm/dd/yyyy" and another data source may use "dd/mm/yyyy". Such differences in formats not only complicate the data processing, but also significantly increase the risk of errors during their analysis. If the schema of the input document does not correspond to the internal schema of documents in the organization, bringing it to the proper format is one of the tasks of data management.

There are solutions that work with the document schema, but for them you need to specify a specific algorithm for processing. Also, the specifics of the data lifecycle in a particular organization can be complex, as data can have different lineage. The better an organization understands the lifecycle and lineage of data, the better it can manage its data [1].

The solution to the described problems is an automated analysis of the schema of incoming documents. For semi-structured data, this schema is the list of input columns, their content, compliance with business needs. To bring the incoming document to a fixed structure, instructions are to be created on how the incoming columns should correspond to the columns of the organization’s internal document schema, the content of each of them is to be reviewed to match the request, if necessary, and the number of rows is to be counted and saved in a commonly accepted format, for example JSON (Fig. 1).

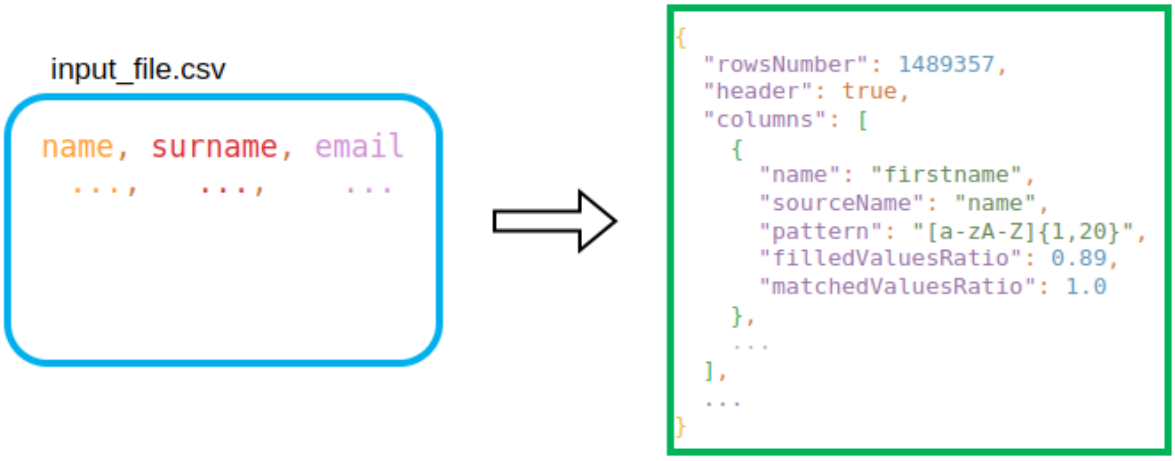

Figure 1. The result of document schema analysis

The result of the analysis will be a recipe that consists of two parts: a description of the schema of the input file, which ensures the preservation of the lineage of the data, and recommendations for transformation. It can be used to make decisions and further process data to align it with the needs of the organization. Automating this process reduces the time required to integrate input data sources and the human risks associated with manual analysis.

The proposed approach has a list of advantages over the analogs [3, 4]. Firstly, it can be easily customized according to the customer's needs and the peculiarities of the business domain. Secondly, it is transparent enough to exclude possible cyber threats connected with data processing in business analytics-oriented applications by large proprietary platforms, where data wrangling and processing is hidden from the customer and can not be traced appropriately. Besides, the developed pipeline imposes no restriction on external datastore connectivity

Possible applications of the developed pipeline range from scientific databases and digital libraries to on-line documentations and electronic commerce. Adoption of the proposed approach will significantly improve the subsequent data validation and querying as well as reduce security and privacy risks, enabling opportunities for efficient backup and disaster recovery.

References:

1. DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd ed.). Technics Publications.
2. S. Abiteboul (2009). Semi-Structured Data. In: Liu L., Özsu M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_799
3. Altair Monarch. URL: https://altair.com/monarch/, accessed 1 May 2024.
4. Alteryx. URL: https://www.alteryx.com/products/alteryx-platform, accessed 1 May 2024.