Structural Preparation of Raw Data Files

Mazhar Hameed

Hasso Plattner Institute
Information Systems Group

University of Potsdam, Germany

Data preparation stands as a cornerstone in the landscape of data science workflows, commanding a significant portion—approximately 80%—of a data scientist's time. The extensive time consumption in data preparation is primarily attributed to the intricate challenge faced by data scientists in devising tailored solutions for downstream tasks. This complexity is further magnified by the inadequate availability of metadata, the often ad-hoc nature of preparation tasks, and the necessity for data scientists to grapple with a diverse range of sophisticated tools, each presenting its unique intricacies and demands for proficiency.

Previous research in data management has traditionally concentrated on preparing the content within columns and rows of a relational table, addressing tasks, such as string disambiguation, date standardization, or numeric value normalization, commonly referred to as data cleaning. This focus assumes a perfectly structured input table. Consequently, the mentioned data cleaning tasks can be effectively applied only after the table has been successfully loaded into the respective data cleaning environment, typically in the later stages of the data processing pipeline.

While current data cleaning tools are well-suited for relational tables, extensive data repositories frequently contain data stored in plain text files, such as CSV files, due to their loose standard. Consequently, these files often exhibit tables with a flexible layout of rows and columns, lacking a relational structure. This flexibility often results in data being distributed across cells in arbitrary positions, typically guided by user-specified formatting guidelines.

Effectively extracting and leveraging these tables in subsequent processing stages necessitates accurate parsing. This thesis emphasizes what we define as the “structure” of a data file—the fundamental characters within a file essential for parsing and comprehending its content. Concentrating on the initial stages of the data preprocessing pipeline, this thesis addresses two crucial aspects: *comprehending* the structural layout of a table within a raw data file and automatically *identifying* and *rectifying* any structural issues that might hinder its parsing. Although these issues may not directly impact the table's content, they pose significant challenges in parsing the table within the file.

Our initial contribution comprises an extensive survey of commercially available data preparation tools. This survey thoroughly examines their distinct features, the lacking features, and the necessity for preliminary data processing despite these tools. The primary goal is to elucidate the current state-of-the-art in data preparation systems while identifying areas for enhancement. Furthermore, the survey explores the encountered challenges in data preprocessing, emphasizing opportunities for future research and improvement.

Next, we propose a novel data preparation pipeline designed for detecting and correcting structural errors. The aim of this pipeline is to assist users at the initial preprocessing stage by ensuring the correct loading of their data into their preferred systems. Our approach begins by introducing SURAGH, an unsupervised system that utilizes a pattern-based method to

identify dominant patterns within a file, independent of external information, such as data types, row structures, or schemata. By identifying deviations from the dominant pattern, it detects *ill-formed* rows. Subsequently, our structure correction system, TASHEEH, gathers the identified ill-formed rows along with dominant patterns and employs a novel pattern transformation algebra to automatically rectify errors. Our pipeline serves as an end-to-end solution, transforming a structurally broken CSV file into a well-formatted one, usually suitable for seamless loading.

Finally, we introduce MORPHER, a user-friendly GUI integrating the functionalities of both SURAGH and TASHEEH. This interface empowers users to access the pipeline's features through visual elements. Our extensive experiments demonstrate the effectiveness of our data preparation systems, requiring no user involvement. Both SURAGH and TASHEEH outperform existing state-of-the-art methods significantly in both precision and recall.