Haiyang ZHU, Dongming HAN, Jiacheng PAN, Yating WEI, Yingchaojie FENG, Luoxuan WENG, Ketian MAO, Yuankai XING, Jianshu LV, Qiucheng WAN, Wei CHEN, 2024. A visual analysis approach for data imputation via multi-party tabular data correlation strategies. *Frontiers of Information Technology & Electronic Engineering*, 25(3):398-414. https://doi.org/10.1631/FITEE.2300480

A visual analysis approach for data imputation via multi-party tabular data correlation strategies

Key words: Data governance; Data incompleteness; Data imputation; Data visualization; Interactive visual analysis

Corresponding author: Wei CHEN
E-mail: hnsyzhy@zju.edu.cn; chenvis@zju.edu.cn
ORCID: Haiyang ZHU, https://orcid.org/0000-0002-4782-5654; Wei CHEN, https://orcid.org/0000-0002-8365-4741

Motivation

1. In large-scale datasets, the abundance of data columns and rows in each table creates challenges in data retrieval. The complexity of data retrieval escalates exponentially with the increasing number of tables, columns, and rows. This poses difficulties in accessing relevant data to fill in the missing values.

2. Due to the heterogeneity of data distribution, types, formats, and structures across various data tables, data columns representing the same attribute can have variations, such as different column names or data distributions. This implies the need for the homogeneity issue of associated data to be effectively addressed, while simple statistical methods may overlook correct correlations.

3. Because of the complexity of data correlations, the workload involved in employing users' background knowledge to verify data becomes substantial and challenging. Therefore, providing guidance for users through the data imputation process is necessary for improving the efficiency.

Main idea

1. We introduce a correlation strategy based on row–column similarity to identify similar data across multi-party tables, thus facilitating a more precise imputation of missing data.

2. We formulate a multi-party tabular data imputation approach by inferring missing data using analogous information from correlated tables, thereby effectively addressing the issue of missing data. We then develop a visual system to support interactive data imputation with our approach.

3. Quantitative and qualitative experiments, as well as user surveys, have demonstrated the effectiveness of our approach in supporting the imputation of missing data based on users' domain knowledge.

System overview

The interactive visual analysis system for missing data imputation discussed in this study consists of three main components: datasets, visual interface, and construction engine.

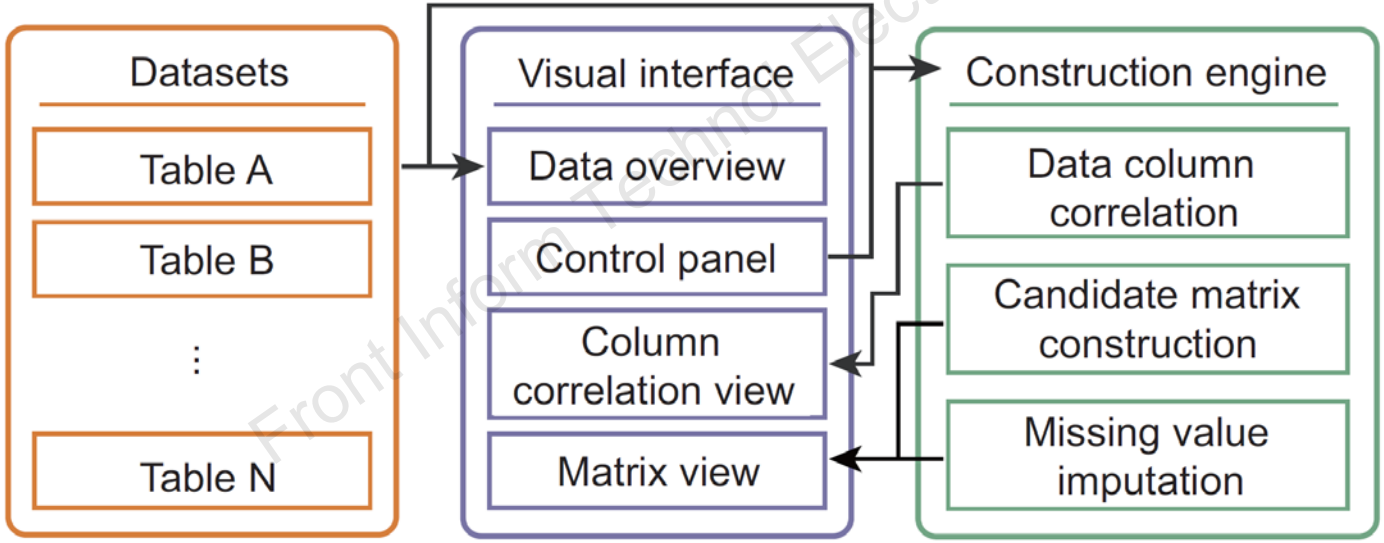

Fig. 1 Overview of our system

Method

The proposed data imputation method consists of three main steps: data column correlation, candidate matrix construction, and missing value imputation.

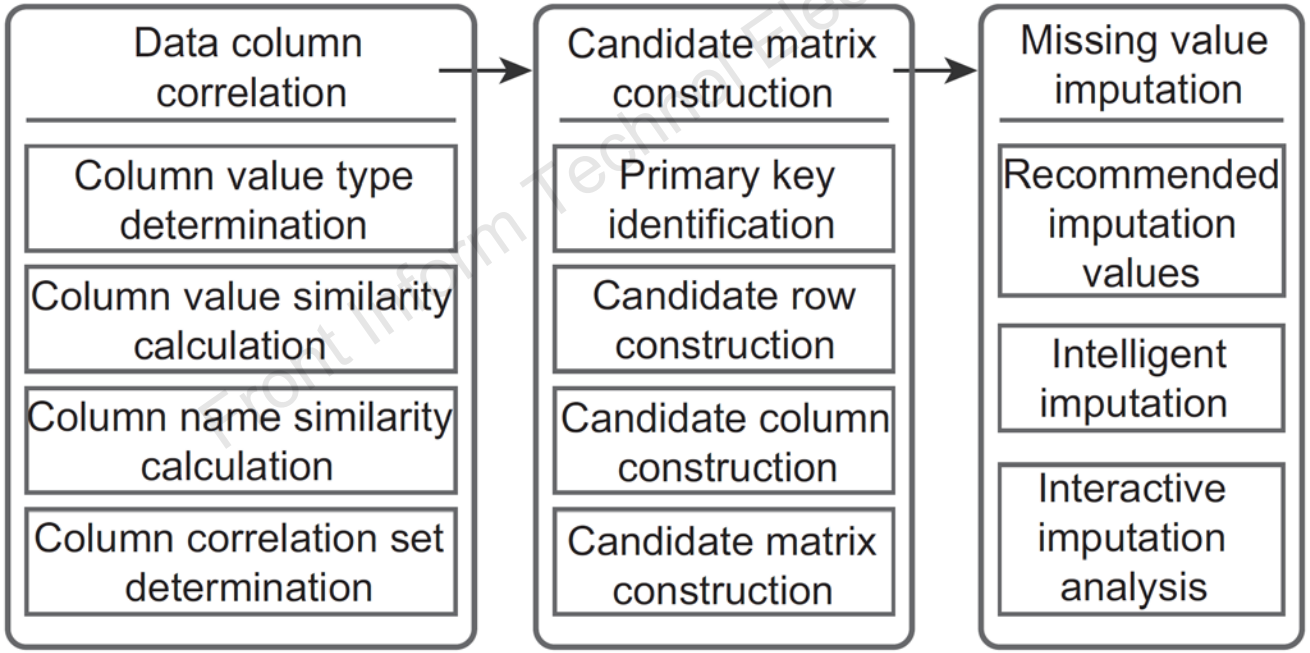

Fig. 2 Overview of our method

Method (Cont'd)

1. Data column correlation

(1) Column value type determination. Our method first assesses the data type of each data column's values cv in all data tables, encompassing four types (identifier, categorical, numerical, and textual).

(2) Column value similarity calculation. Our method computes similarities among column values cv that belong to different data tables but share the same data type, resulting in a matrix of column value similarities M_{cv}.

(3) Column name similarity calculation. To construct the column name similarity matrix M_{cn}, the textual data similarity calculation method discussed previously is employed for the column names.

(4) Column correlation set determination. For data columns from different data tables, our method combines the column value similarity matrix M_{cv} and the column name similarity matrix M_{cn} to calculate the overall similarity matrix M_c.

Method (Cont'd)

2. Candidate matrix construction

(1) Primary key identification. We locate the primary key v in the row containing missing data. By locating the primary key in all rows with missing data, we can obtain the set of primary key values K.

(2) Candidate row construction. For each key value $v_i \in K$, we search for data rows in different data tables that contain v_i and retrieve the corresponding row data to construct candidate rows.

(3) Candidate column construction. Based on the column correlation set C^R, we extract and merge the correlated columns in the candidate rows.

Method (Cont'd)

(4) Candidate matrix construction. Our method further merges the new data table generated in step 3. The column correlation set C^R contains multiple groups of correlated columns. This means that users can establish more accurate column correlation information and decide whether to update the entire engine.

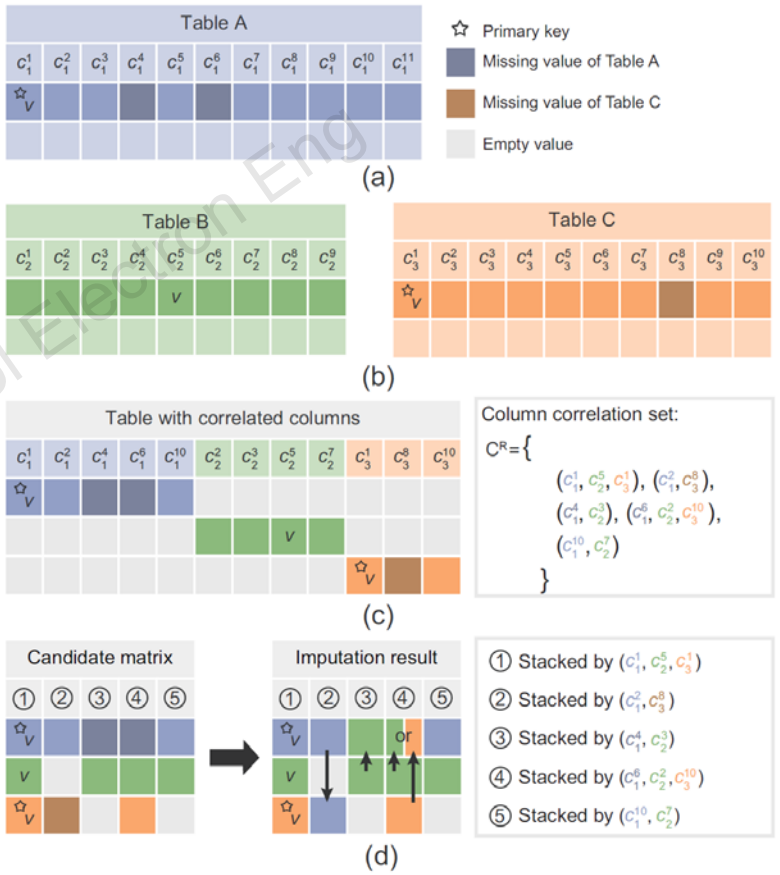

Fig. 3 Candidate matrix construction process: (a) primary key identification; (b) candidate row construction; (c) candidate column construction; (d) candidate matrix construction and data imputation with recommended content

Method (Cont’d)

3. Missing value imputation

(1) Recommended imputation values. The resulting candidate matrix is the recommended outcome for filling in missing data. Each column in the candidate matrix is formed by stacking the candidate row information from several correlated columns.

(2) Intelligent imputation. For multiple candidate matrices, this method provides an intelligent imputation strategy. It chooses the matrix with the highest overall similarity in candidate columns as the imputation-recommended value. Users can determine whether to keep the automatically recommended values through interactive analysis.

(3) Interactive imputation and analysis. Users can determine the imputation of missing content based on the source of the candidate matrix construction.

Method (Cont'd)

4. Visual interface

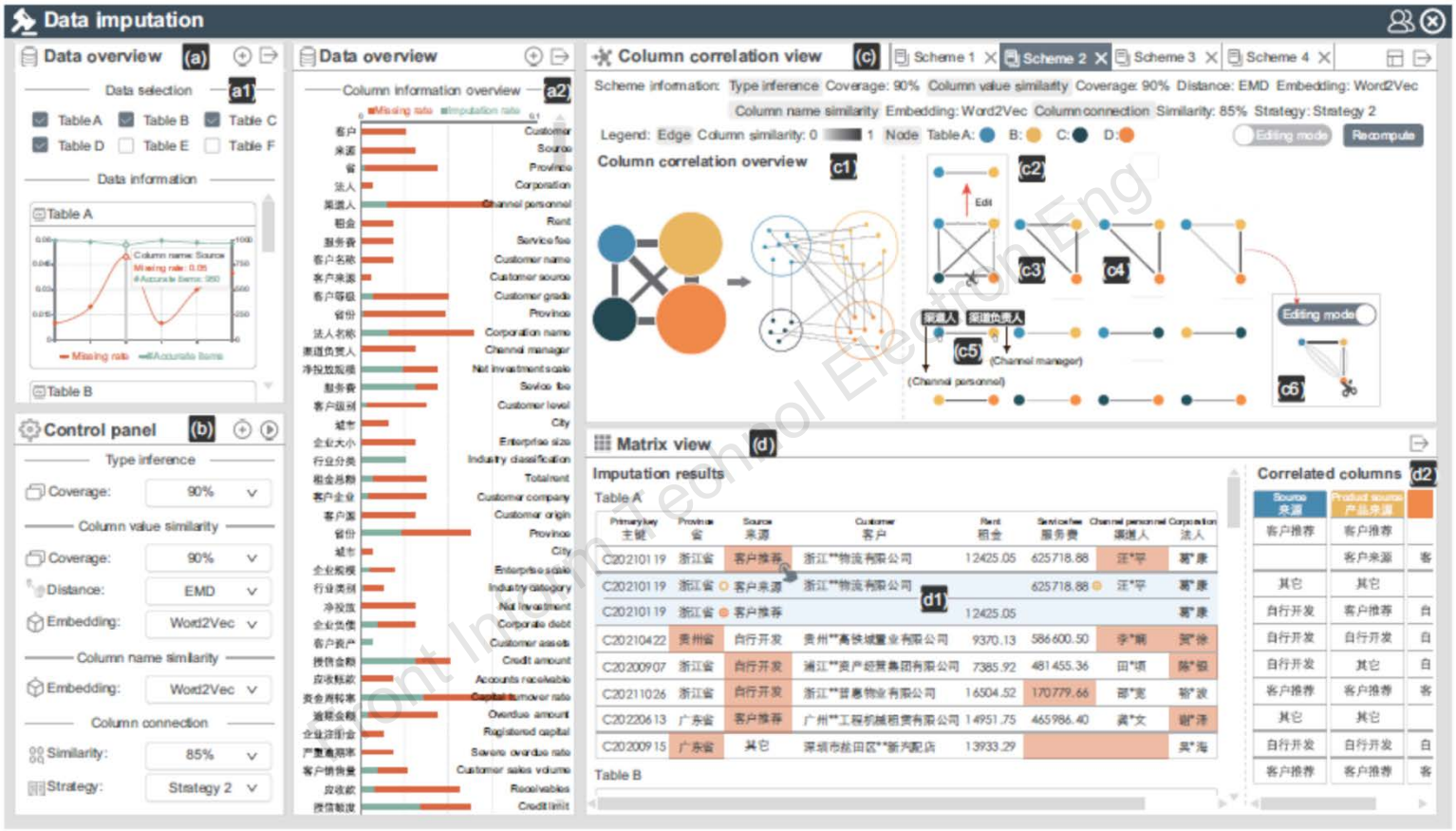

Fig. 4 Visual interface: (a) data overview, used for data quality investigation and offering a column information overview; (b) control panel, allowing users to adjust the methods, strategies, and relevant parameters for column correlations; (c) column correlation view, providing insights into data column correlations and enabling user-driven interactive modification and data relation reconstruction by the user; (d) matrix view, presenting candidate matrix information, recommended imputation values for missing data, and their sources. It also supports interactive selection of recommended imputation values for filling in missing data

Method (Cont’d)

(1) Data overview (Figs. 4a1 and 4a2) helps users understand the characteristics of the data and identify missing data.

(2) Control panel (Fig. 4b) is designed for configuring the column value type determination, column value similarity calculation, and column name similarity calculation.

(3) Column correlation view (Fig. 4c) helps users understand the correlations between columns in different data tables. It displays results for different schemes and supports switching among them. It also enables interactive modification and reconstruction of data relations by the user.

(4) Matrix view (Fig. 4d) displays recommended values for data imputation along with information in the candidate matrix and sources of the recommended imputation values. Users can select recommended imputation values to fill in missing cells (highlighted in red).

Conclusions

1. This paper presents a data imputation method based on a multi-party table data association strategy and constructs an efficient interactive data imputation visual analysis system to assist data governance professionals in addressing the accuracy and effectiveness issues prevailing in relation to missing values.
2. The system uses existing data information from multi-dimensional data tables for imputation. It builds a candidate matrix from the multi-dimensional data tables, associates column information from different data tables, and leverages the background knowledge of data governance professionals to interactively analyze, comprehend, and determine recommended imputation values for missing data in the candidate matrix.
3. The proposed method is validated using a dataset from a large-scale supply-chain enterprise and is compared with existing mainstream data imputation methods through comparative experiments and user surveys. The effectiveness and practicality of the method are quantitatively evaluated and validated.

Haiyang ZHU, a Ph.D. candidate and professor-level senior engineer at the School of Computer Science and Technology, Zhejiang University. He is recognized as a leading talent in scientific and technological innovation under the "Ten Thousand Talents Program" in Zhejiang Province (2022). His primary research areas include digital transformation, data elements, data governance, and big data visual analysis. Serving as the project leader, he has undertaken three provincial key R&D projects, including the Key R&D “Pioneer” Tackling Plan Program of Zhejiang Province, and has participated in two National Key R&D Projects. He has published 21 research papers in SCI/EI/Core Journals, obtained 19 authorized national invention patents, and contributed to the formulation of six technical standards in the information field. He has been awarded one second-class prize for Provincial and Ministerial-Level Scientific and Technological Progress.

Wei CHEN, a professor in State Key Lab of CAD & CG at Zhejiang University, China. His current research interests include visualization, visual analytics, and AI. He has performed research in visualization and visual analysis and published more than 90 IEEE/ACM Transactions and IEEE VIS papers. His Chinese books on visualization are the unique books on visualization in China. He actively served in many leading conferences and journals, like IEEE PacificVIS steering committee, ChinaVIS steering committee, paper co-chairs of IEEE VIS, IEEE PacificVIS, IEEE LDAV and ACM SIGGRAPH Asia VisSym. He is an associate editor of *IEEE TVCG*, *IEEE TBG*, *ACM TIST*, *IEEE T-SMC-S*, *IEEE TIV*, *IEEE CG&A*, *FCS*, and *JOV*.