Record Fusion via Inference and Data Augmentation

ALIREZA HEIDARI, Department of CS, University of Waterloo, Canada
GEORGE MICHALOPOULOS, Department of CS, University of Waterloo, Canada
IHAB F. ILYAS, Department of CS, University of Waterloo, Canada
THEODOROS REKATSINAS, Department of CS, ETH, Switzerland

PROBLEM STATEMENT

Record fusion involves merging duplicate records from different sources into a single, unified record. It helps to improve data quality, reduce redundancy, and enable more accurate analysis and decision-making. However, the process can be challenging due to inconsistencies in data, spelling variations, missing data, or intentional duplications. Record fusion is a critical step in data management that helps organizations make informed decisions, reduce costs, and improve efficiency. Ultimately, record fusion enables organizations to better leverage the value of their data and gain a competitive advantage.

METHODS

The proposed learning framework for record fusion uses a weak supervision approach to automatically combine related data, such as integrity constraints, data rules, quantitative statistics, and source information. The approach addresses several technical challenges, such as the need for an expressive model to capture all data characteristics, the difficulty of gathering enough labeled examples, and the limited information available for designing a reliable ML approach. To overcome these challenges, the approach generates additional training data automatically, and the model learns from partial and noisy estimates of the correct values. The goal is to infer the correct values with an expressive model that captures all data context features, and the approach uses an iterative mechanism to improve predictions over time.

RESULTS

The proposed method for record fusion achieves an average precision of around 98% when source information is available and around 94% without source information. This is a significant improvement over previous approaches and demonstrates the effectiveness of the method in merging data records. The approach also improves the precision compared to other data fusion and entity consolidation methods by an average of around 20/45 precision points with/without source information. Additionally, the data augmentation method used in this approach improves previous approaches by an average of around 10 precision points.

SIGNIFICANCE

The proposed machine learning framework offers an effective solution to the problem of merging data records that combines two well-known problems: data fusion and golden record. By using rich data representation models, data augmentation techniques, and iterative model applications, the approach achieves a high level of accuracy and outperforms previous methods. These results demonstrate the effectiveness and significance of the proposed method, offering an efficient solution to the problem of merging data records and improving the quality of data analysis and decision-making.

Keywords: Record Fusion, Data Cleaning, Data Augmentation, Data Inference

Authors' addresses: Alireza Heidari, a5heidar@uwaterloo.ca, Department of CS, University of Waterloo, Canada; George Michalopoulos, gmichalo@uwaterloo.ca, Department of CS, University of Waterloo, Canada; Ihab F. Ilyas, ilyas@uwaterloo.ca, Department of CS, University of Waterloo, Canada; Theodoros Rekatsinas, a5heidar@uwaterloo.ca, Department of CS, ETH, Switzerland.

Research Article