SHORT-PAPER

CoClean: Collaborative Data Cleaning

Authors: Mashaal Musleh, Mourad Ouzzani, Nan Tang, AnHai Doan | Authors Info & Claims

SIGMOD '20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data
Pages 2757 - 2760 https://doi.org/10.1145/3318464.3384698

Published: 31 May 2020 Publication History Check for updates

11 525

Get Access

Feedback

Abstract

High quality data is crucial for many applications but real-life data is often dirty. Unfortunately, automated solutions are often not trustable and are thus seldom employed in practice. In real-world scenarios, it is often necessary to resort to manual cleaning for obtaining pristine data. Existing human-in-the-loop solutions, such as Trifacta and OpenRefine, typically involve a single user. This is often error-prone, limited to a single-person expertise, and cannot scale with the ever growing volume, variety and veracity of data.

We propose a crowd-in-the-loop cleaning system, called CoClean, built on top of

one to share data represented as a dataframe with other users. CDF is responsible for synchronizing and aggregating annotations obtained from

(or a subset of it) to different users. (2)Supporting both lay and power users: lay users can use a GUI for direct manual cleaning of the data, while power users can work on the assigned data through a Jupyter Notebook where they can write

which can make the life of users easier for manual cleaning. (4)Collaboration Modes: CoClean supports two modes: blind-on(no user can see the annotations from others) and blind-off.

References

[1] Z. Abedjan, X. Chu, D. Deng, R. C. Fernandez, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang. Detecting data errors: Where are we and what needs to be done? PVLDB, 9(12):993--1004, 2016.

Digital Library | Google Scholar

[2] M. Mahdavi, Z. Abedjan, R. C. Fernandez, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang. Raha: A configuration-free error detection system. In SIGMOD, pages 865--882, 2019.

Digital Library | Google Scholar

[3] A. A. Qahtan, A. Elmagarmid, R. Castro Fernandez, M. Ouzzani, and N. Tang. Fahes: A robust disguised missing values detector. In ACM SIGKDD, 2018.

Digital Library | Google Scholar

[4] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. Holoclean: Holistic data repairs with probabilistic inference. PVLDB, 10(11):1190--1201, 2017.

Digital Library | Google Scholar

Cited By

View all

Information Science and Systems. 10.1007/s13755-024-00295-6. **12**:1. Online publication date: 5-Jul-2024.

Yang H, Zhang G, Su Y and Guo N. (2024). A power fusion data cleaning method based on exponential moving average and cosine similarity algorithms. 2024 IEEE 10th International Conference on Edge Computing and Scalable Cloud (EdgeCom). 10.1109/EdgeCom62867.2024.00012. (25-30). Online publication date: 28-Jun-2024.

https://doi.org/10.1109/EdgeCom62867.2024.00012

Perini M and Nikolic M. (2024). In-Database Data Imputation. Proceedings of the ACM on Management of Data. 10.1145/3639326. **2**:1. (1-27). Online publication date: 26-Mar-2024.

https://dl.acm.org/doi/10.1145/3639326

Show More Cited By

Index Terms

CoClean: Collaborative Data Cleaning

Information systems

Data management systems

Information integration

Data cleaning

Recommendations

Data Cleaning: Overview and Emerging Challenges

in inaccurate analytics and unreliable decisions. Over the past few years, there has been a surge of interest from...

Read More

Consolidation of the research information improves the quality of data integration, reducing duplicates between…

Read More

Usability of Visual Data Profiling in Data Cleaning and Transformation

On the Move to Meaningful Internet Systems. OTM 2017 Conferences

Abstract
This paper proposes an approach for using visual data profiling in tabular data cleaning and transformation…

Read More

Comments

DL Comment Policy

Comments should be relevant to the contents of this article, (sign in required).

Got it

0 Comments

Share

Best Newest Oldest

Nothing in this discussion yet.

Download PDF

Books

Proceedings

SIGs

Conferences

Collections

People

Subscription Information

Author Guidelines

Using ACM Digital Library

All Holdings within the ACM Digital Library

ACM Computing Classification System

Accessibility Statement

Join

Join ACM

Join SIGs

Subscribe to Publications

Institutions and Libraries

Connect

Contact us via email

ACM on Facebook

ACM DL on X

ACM on Linkedin

Send Feedback

Submit a Bug Report