POSTER: Modeling Provenance and Understanding Reproducibility for OpenRefine Data Cleaning Workflows

Timothy McPhillips Lan Li Nikolaus Parulian Bertram Ludäscher
School of Information Sciences, University of Illinois at Urbana-Champaign
{tmcphill,lanl2,nnp2,ludaesch}@illinois.edu

1 INTRODUCTION

Preparation of data sets for analysis is a critical component of research in many disciplines. Recording the steps taken to clean data sets consequently is essential to making such research transparent and results reproducible. OpenRefine is a tool for interactively cleaning data sets via a spreadsheet-like interface [VDW13]. OpenRefine records the sequence of operations carried out and changes made to a data set. It uses these records to provide an undo/redo capability that allows a user to revisit the past state of the data set at any point following the initial data import step. OpenRefine further enables the user to export sequences of these recorded operations as recipes that can be saved and applied later to different data sets. However, because such exported recipes do not include edits made manually to individual cells, a recipe cannot represent the end-to-end data preparation workflow. Here we investigate how the complete operation history recorded by OpenRefine can be used both to facilitate reproduction of complete, real-world data cleaning workflows and to support queries and visualizations of the provenance of cleaned data sets.

2 AIMS

Describe OR's history capabilities using concepts and terminology familiar to members of the provenance research community. Employ the operation history captured by OR to satisfy queries about the provenance of the cleaned data sets. Identify provenance queries that can be supported by OR's native data model and operation history. Experiment with accessing and using information in the history records beyond what OR itself allows one to view or export as recipes. Represent overall data cleaning workflows carried out in OR in YesWorkflow, and extend the YW data model to enable us to take advantage of YW visualization and query support in the contect of data cleaning workflows. Identify apparent limitations on kinds of provenance queries that can be supported given OR's native data model and operation history.

3 TOOLS

OpenRefine 3.1 distribution installed in a Linux environment. OpenRefine REST API and Python client library, with custom extensions, for automating operation of OpenRefine YesWorkflow toolkit for modeling data cleaning workflow and representing OR operation history in queryable form. XSB Prolog for expressing and performing Datalog-style graph and provenance queries. GraphViz for rendering visualizations of query results. GitHub for sharing research artifacts between co-authors and with research community. Ansible, Vagrant, and Docker for making research environment reproducible across coauthors' computers and for enabling other researchers to repeat our experiments on their own computers. Whole Tale and MyBinder for enabling others to reproduce our results without installing software on their own computers.

4 EXAMPLE PROVENANCE QUESTIONS

5 RESULTS

We demonstrate that under certain conditions complete data cleaning workflows carried out within OR can be repeated fully automatically in a different instance of OR. We shown that this requires using information that is recorded by OR for its undo/redo feature, but that is not exportable from OR via recipes or its native HTTP API. We show that key queries of the provenance of cleaned data sets, and of particular columns, rows, and cells in the final data set, can be satisfied using the information captured by OR for its undo/redo feature. Will illustrate the usefulness of YesWorkflow-style workflow diagrams for making data cleaning workflows transparent and easy to review, and for rendering portions of the overall workflow to represent the result of provenance queries.

6 FUTURE WORK

REFERENCES

[VDW13] R. Verborgh and M. De Wilde. *Using OpenRefine: The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web.* Community experience distilled. Packt Publishing, Birmingham Mumbai, 2013. OCLC: ocn892971028.