Poster Abstract: 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 many research projects. When data to be used in research was originally entered manually or is taken from multiple sources, the data generally must be cleaned prior to use. OpenRefine (OR) is a commonly used tool both interactively and semi-automatically cleaning data sets. OR records the sequence of operations carried out during data cleaning and allows the user to review this history through the user interface. For each data cleaning operation carried out OR creates a history entry, representing metadata about the operation; and a change object that records the actual changes made to the data set as a result of the operation. The operation history and changes recorded by OpenRefine are sufficient to support for full undo/redo capability, and OR provides a Undo/Redo feature that allows a user to revisit the state of the data set at any point during the data cleaning following the initial data import step. OR operations are considered generalizable if they can in principle affect more than one cell of a data set. OR enables the history entries corresponding to generalizable operations to be exported as OR recipes that can then be applied to other data sets through OpenRefine. An OR recipe can be a single operation, or a sequence of operations that can be carried out automatically (and so may be treated as a reusable data-cleaning macro). [VDW13]

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.