An architecture for end-user informed data preparation

Data systems:

Value Added Information & Technology

Data systems (VADA) and its value. Dr Norman Paton, Professor of Computer Science at the University of Manchester, are carrying out research into Value Added Data systems (VADA) and exploring various techniques for automating the creation of a data wrangling process, as a whole a process that requires both significant skill and time.

DATA PREPARATION

A number of data preparation approaches are currently in widespread use. These tend to fall into three groups: those that involve programming the solution; those that develop workflows that extract, transform and load the data into analysis platforms; and those that develop transformations using tabular representations of the data. These data preparation tools usually offer components that support similar tasks, such as combining data sets and reformattting columns, but they differ in how these can be expressed by data scientists. Even with the support of these tools, data scientists are required to retain fine-grained control over each aspect of the process. While this is appropriate in some circumstances, the costs are high and can be prohibitive.

AUTOMATING THE CREATION OF DATA PREPARATION PROCESSES

As the myriad of data sources within organisations and in the public domain continues to grow, so does the need for cost-effective, scalable and principled techniques for integrating and cleaning big data.

DATA PREPARATION PROCESSES

The research team have observed that automating data wrangling in this way involves being able to identify the steps in the data preparation process that can be automated using available evidence and feedback. They are then able to develop components for these steps that take all of the available evidence and feedback into account. This leads to the development of an overall architecture that allows the user to provide the evidence, view the results, and then provide feedback.

EVALUATION OF VADA

The researchers have carried out an empirical evaluation of VADA using a case study involving real world Web extracted data from real-estate sources.

The resulting automated solution may not be suitable or even correct and the user can provide feedback as to the correctness or suitability of the result. In the light of the new evidence, the system can then act on this feedback and automatically generate a revised approach to populating the target. This process is repeated until the result is deemed fit for purpose.

The research team have observed that automating data wrangling in this way involves being able to identify the steps in the data preparation process that can be automated using available evidence and feedback. They are then able to develop components for these steps that take all of the available evidence and feedback into account. This leads to the development of an overall architecture that allows the user to provide the evidence, view the results, and then provide feedback.

www.researchoutreach.org www.researchoutreach.org
In the VADA data preparation process, users need only specify what they need, not how to produce it. Considerable improvements with f-scores for precision and recall were demonstrated in examples. Using the f-score to combine the results with and without the data wrangling process, researchers ran a five-step automated process. Potential value through analysis was translated into practice.

**VADA BENEFITS**

These encouraging results show that with the VADA user interface, data scientists can obtain cleaned and integrated data from multiple data sets, having provided only a target schema and associated data context. The process requires only modest effort when compared to the intense manual involvement that is currently required. For example, the automatically produced wrangling process combines data sets, reformats inconsistent attribute values, and resolves certain inconsistencies. Manually crafting queries and rules to carry out such tasks requires significant skill and effort.

VADA also takes user preferences and user feedback into account. By capturing preferences through the user context, data can be selected in ways that trade-off the accuracy, consistency and relevance of the wrangling result. Furthermore, the automation means that many alternative candidate data products can be produced. The evaluation of the user context demonstrates how feedback can impact on the utility of the results. Moreover, this approach can wrangle hundreds of sources in minutes. This scale of data preparation would likely take many days of manual effort.

**SPIN-OUT COMPANIES**

There are two spin-out companies that translate different aspects of this work into practice.

Firstly, The Data Value Factory (https://datavaluefactory.com) which offers automated data preparation. The Data Preparer system builds on the research team’s experience and provides the first declarative data wrangling software product. A free trial version of the Data Preparer wrangling platform is available for download.

Secondly, DeepReason.ai (https://deepreason.ai) enables ‘Knowledge-Find’ AI Solutions. It uses a Knowledge Graph platform that operates state-of-the-art AI technology to deliver end-to-end AI solutions to Fortune 500 companies in finance, logistics, manufacturing, and engineering.

**FUTURE WORK**

The University of Manchester research team are exploring how they can combine automated data preparation with discovery of relevant data sets in data lakes. In addition, automating data preparation creates opportunities to explore how the introduction of bias during data preparation can detected and reduced.

There is a need for cost-effective, scalable and principled techniques for integrating and cleaning big data, creating data sets suitable for downstream analysis and the UK open government data portal. These data sets are produced by a large number of independent publishers, so inconsistencies need to be resolved in order to maximise their potential value through analysis. The researchers ran a five-step automated wrangling process. They then compared the results with and without data context. Using the faceted to combine precision and recall, examples showed considerable improvements with f-scores rising from around 0.5 to 0.8 when data context was used as evidence throughout the wrangling process.

**References**


**Research Objectives**

Research at the Information Management Group at the University of Manchester focuses on distributed information management for challenging environments and applications.

**Detail**

Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK

**Bio**

Norman Paton: Norman has been a Professor of Computer Science at Manchester University since 2000, and is now a Founder/Director at The Data Value Factory, working to commercialise techniques on cost-effective data preparation. His research has focused on distributed information management, including applications in the life sciences.

Nikolaos Konstantinou: Nikos has been a Research Fellow at Manchester since 2015, prior to which he held a variety of technical management and research roles in Greece. He is a Founder/Director at The Data Value Factory, working to bring innovative data preparation techniques to market.

**Collaborators**

• Edward Abel, Alex Bogatu, Martin Koehler, Lacramioara Mazilu, Alvaro Fernandes, John Keane, School of Computer Science, University of Manchester
• Luigi Bellomarini, Emanuel Sallinger, Georg Gottlob, Department of Computer Science, University of Oxford
• Cristina Civili, Leonid Libkin, School of Informatics, University of Edinburgh

**Personal Response**

We worked over a considerable period with life science researchers, who often need to combine experimental results with existing information about an organism. In these collaborations, it became clear that the cost of preparing the data for analysis was a significant barrier to progress.

**Behind the Research**

**Norman Paton**

E: norman.paton@manchester.ac.uk T: +44 1612756491 W: https://www.research.manchester.ac.uk/portal/norman.paton.html

**Nikolaos Konstantinou**

E: nikolaos.konstantinou@manchester.ac.uk T: +44 1612756583 W: https://personalpages.manchester.ac.uk/staff/nikolaos.konstantinou

**What initially sparked your interest in automating data wrangling?**

We worked over a considerable period with life science researchers, who often need to combine experimental results with existing information about an organism. In these collaborations, it became clear that the cost of preparing the data for analysis was a significant barrier to progress.