Systematically Corrupting Data to Assess Data Linkage Quality

Ahmad Alsadeeqi[1], Alasdair Gray[1], Özgür Akgün[2], Tom Dalton[2], Peter Christen[3]

[1] Heriot-Watt University, Edinburgh, UK
[2] School of Computer Science, University of St Andrews, UK
{ozgur.akgun, tsd4}@st-andrews.ac.uk
[3] The Australian National University, Canberra, Australia

Various algorithms have been developed to automatically link historical records based on a variety of string matching techniques. These generate an assessment of how likely two records are to be the similar. However, it remains unclear how to assess the quality of the linkages computed due to the absence of absolute knowledge of the correct linkage of real historical records - the ground truth. The creation of synthetically generated datasets for which the ground truth linkage is known to help with the assessment of linkage algorithms but the data generated is commonly too clean to be representative of historical records. We are interested in assessing record linkage algorithms under different data quality scenarios, e.g. with errors typically introduced by a transcription process or where books can be nibbled by mice. We are developing a data corrupting model that injects corruptions into datasets based on given corruption methods and probabilities. We have classified different forms of corruptions found in historical records into four types based on the effect scope of the corruption. Those types are character level (e.g. an 'f' is represented as an 's' - OCR Corruptions), attribute level (e.g. gender swap - male changed to female due to false entry), record level (e.g. missing records due to different reasons like loss of certificate), and group of records level (e.g. lost parish records in fire). This will give us the ability to evaluate record linkage algorithms over synthetically generated datasets with known ground truth and with data corruptions matching a given profile. In this paper, we describe in detail these four types of corruptions and corresponding examples.

This work was presented at the following workshop.

Venue	ADRN 2017 - The UK Administrative Data Research Network Annual Research Conference
Location	The Royal College of Surgeons of Edinburgh
Date	June 1-2, 2017
URL	http://www.adrn2017.net/