Record Linking Using Metric Space Similarity Search

Alan Dearle, Graham Kirby, Özgür Akgün, Tom Dalton

School of Computer Science, University of St Andrews, UK
{alan.dearle, graham.kirby, ozgur.akgun, tsd4}@st-andrews.ac.uk

Record linking often employs blocking to reduce the computational complexity of full pairwise comparison. A key is formed from a subset of record attributes. Those records with the same key values are blocked together for detailed comparison. Use of a single blocking key fails to detect many true matches if records contain missing values or errors, since only those records with the same key values are compared. To address missing values, it is common to repeat the matching process using multiple blocking keys, to match records that are identical in a subset of the fields. The presence of erroneous values may be addressed by blocking using key values mapped to a canonical form (e.g. Soundex). However, this does not address other problems such as single digit transcription errors in dates. Blocking is used to categorise records that are candidate matches, in preparation for a pairwise comparison phase which may use various distance metrics, depending on the domain of the values being compared. Each blocking process defines a partition of records. The comparison operations are only applied to pairs of records within the same category. In some contexts, it may be useful to have flexible control over the precision/recall trade-off, depending on the intended use for the matched data, and the degree of conservatism required of the identified links. With blocking, this flexibility is limited by the number of sensible blocking keys that can be identified. In this talk, we describe experiments with a technique based on similarity searching over metric spaces, which appears to offer greater flexibility, and describe some preliminary results using an historic Scottish dataset.

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/