International Journal of Population Data Science

Journal Website: www.ijpds.org

High quality linkage using Multibit Trees for privacy-preserving blocking

Brown, Adrian[1], Borgs, Christian[2], Randall, Sean[1], and Schnell, Rainer[2]

[1]Curtin University

[2]University of Duisburg-Essen

[3]Massachusetts Department of Public Health

Objectives

As privacy-preserving record linkage (PPRL) emerges as a method for linking sensitive data, efficient blocking techniques that help maintain high levels of linkage quality are required. This research looks at the use of a Q-gram Fingerprinting blocking technique, with Multibit Trees, and applies this method to real-world datasets.

Approach

Data comprised ten years of hospital and mortality records from several Australian states, totalling over 25 million records. Each record contained a linkage key, as defined by the jurisdiction, which was used to assess quality (i.e. used as a 'gold standard'). Different parameter sets were defined for the linkage tests with a privacy-preserved file created for each parameter set. The files contained jurisdictional linkage key and a Cryptographic Long-term Key (the CLK is a Bloom filter comprising all fields in the parameter set).

Each file was run through an implementation of the Q-gram Fingerprinting blocking algorithm as a deduplication technique, using different similarity thresholds. The quality metrics of precision, recall and f-measure were calculated.

Results

Resultant quality varied for each parameter set. Adding suburb and postcode reduced the linkage quality. The best parameter set returned an F-measure of 0.951. In general, precision was high in all settings, but recall fell as more fields were added to the CLK. We will report details for all parameter settings and their corresponding results.

Conclusion

The Q-gram Fingerprinting blocking technique shows promise for maintaining high quality linkage in reasonable time. Determining which fields to include in the CLK for the linkage of specific datasets is important to maximise linkage quality, as well as selecting optimal similarity thresholds. Developing new technology is important for progressing the implementation of PPRL in real-world settings.

*Corresponding Author:
Email Address: adrian.brown@curtin.edu.au (A. Brown)

http://dx.doi.org/10.23889/ijpds.v1i1.149