A Machine Learning Approach for Nominative Record Linkage in Chinese Historical Databases

Yue Yu

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

yue.yu@connect.ust.hk

Cameron Campbell

Division of Social Science, The Hong Kong University of Science and Technology

Abstract

We introduce a generic machine learning-based pipeline for nominative linkage of records within and across large-scale Chinese historical datasets. The pipeline addresses key challenges, including character variations, incomplete data, and scalability issues specific to historical datasets in which names and other attributes are recorded with Chinese characters, not just in China, but potentially in Korea, Japan and Vietnam. Our approach integrates stroke-based character embeddings for efficient blocking, supervised classification with active learning for record matching, and graph-based clustering for final linkage. We demonstrate the effectiveness of this pipeline using the career records of officials in the China Government Employee Database-Qing *Jinshenlu* (CGED-Q JSL) as a test case. We achieve improved linkage quality compared to standard probabilistic methods, with substantially longer linked sequences of career records and fewer aberrant transitions. To validate the generalizability, we also successfully apply the pipeline to another database and a cross-database linkage task. By minimizing the need for manual tuning, our pipeline offers a more accessible and effective solution for Chinese historical data linkage.

Acknowledgements

This research was supported by HK RGC Areas of Excellence AoE/B-704/22-R (Chen Zhiwu PI).