International Journal of Population Data Science

Journal Website: www.ijpds.org

Advanced methods for linking complex historical birth, death, marriage and census data

Christen, Peter^{1*}

¹Research School of Computer Science, The Australian National University

Objective

Recent years have seen the development of novel techniques for linking complex types of data that contain records about different types of entities, for example bibliographic databases with records about authors, publications, and venues. Advanced approaches have been devised to link individuals and groups of records. These approaches exploit both the similarities between record attributes as well as the relationships between entities. Rather than linking records about different types of entities, in this work we study the novel problem of linking records where the same entity can have different roles and where these roles can change over time.

Approach

We specifically develop novel techniques for linking historical birth, death, marriage, and census certificates with the aim to reconstruct the population covered by these certificates over a period of time. Our techniques make use of constraints that consider roles, relationships, as well as time. Our first technique links certificates based on the specific roles of their individuals, and greedily selects pairs of certificates with the highest overall similarity while also considering 1-to-1 and 1-to-many linkage constraints. Our second hybrid technique combines graph, group, and temporal linkage, and also considers relationship information between individuals and groups. We compare these techniques with state-of-the-art group, collective, and graph-based linkage approaches.

Results

We evaluate our proposed techniques on real Scottish data from 1861 to 1901 that cover the population of the Isle of Skye. In total, these data sets contain 119,042 certificates for 234,365 individuals. As ground truth we have a set of life-segments of

records manually linked by domain experts. Our results indicate that even advanced techniques have difficulty in achieving high linkage quality compared to careful manual linkage. Two reasons for this are the very small name pool in our data and the changing nature of people's personal details over time. Both our proposed techniques, however, significantly outperform traditional pair-wise attribute similarity and group linkage approaches, with the greedy role-based technique achieving better results than the hybrid technique.

Conclusion

Our experiments on real data show that even with advanced linkage techniques that employ group, graph, relationship, and temporal approaches it is challenging to achieve high quality links from complex data such as birth, death, marriage and census certificates that span several decades. As future work we will improve all steps of our techniques with the goal of developing highly accurate, scalable, and automatic techniques for linking large-scale complex population databases.

*Corresponding Author:

Email Address: peter.christen@anu.edu.au (P. Christen)