International Journal of Population Data Science

Journal Website: www.ijpds.org

Incremental Interactive Record Linkage using Human Intelligence Tasks (HITs)

Kum, Hye-Chung[1*]

[1]Texas A&M University

Objectives

When analyzing population data, there is a need to link data about organizations. One challenge in linking organization level data is that unlike a person, there can be many definitions for an entity. For example, for hospitals, depending on the dataset, an entity might represent any one of the following similar but different semantic types:(1) physical units, (2) billing units (3) legal units, (5) licensed units, or (5) reporting units. How these different entities relate to each other can be complex such as one billing unit can span across many physical units or multiple billing units can exist for one physical unit. Thus, linking organization level data requires human involvement to sort through these issues in heterogeneous data sources to make informed decisions on the messy data. We design and evaluate a general framework for a hybrid Human-Machine process for ongoing integration and cleaning of hospital level data when no common identifiers exist such that we highlight the decisions that need human judgement and document and track the full processes to ensure reproducibility. Such ongoing integration is often called incremental record linkage (RL).

Approach

Accurate linkage in big data requires well-defined tasks that need automatic or human processing. In the human computer interaction (HCI) field, Human Intelligence Tasks (HITs) are defined as micro tasks requiring human judgment and are often used in designing crowdsourcing systems. We designed HITs for linking organization level data and embed them into automatic deterministic linkage algorithms that supports interactive stepwise RL. The hybrid system is a framework for reproducible incremental RL.

Results

We illustrate this framework by integrating four databases of hospitals in Texas from 2008 to 2014(N=664). The IDs used in the databases are the Texas Provider ID, the National Provider ID, the Medicare ID, and the Facility ID. We link the databases using provider name, including dba (i.e., doing business as), addresses, and phone numbers. Similarities in hospital names and addresses and the dynamic nature of hospital attributes over time make it impossible to build a fully automated linkage system for hospitals. Using our system to iteratively standardize and clean the data, we linked the hospitals with 100% precision using HITs that required confirming 79 approximate linkages and manually linking 28 hospitals.

Conclusion

Effective software that can support the interactive and iterative process of RL with well-designed HITs can streamline the linkage processes to support high quality replicable research using big data.

*Corresponding Author:
Email Address: kum@tamu.edu (H. Kum)

http://dx.doi.org/10.23889/ijpds.v1i1.231