Short summary of activities of NCRN-CORNELL node

2012-07-05
Lars Vilhuber, John Abowd, Ping Li, William Block

The NCRN-Cornell node (Node) is working on Comprehensive Extensible Data Documentation and Access Repository (CED²AR, previously referred to as CCBMR), as well as on applying modern machine learning techniques for Census-related applications. Our main website (www.ncrn.cornell.edu) is live, but is still being expanded. (a) We have completed high level functional diagram initial technical diagram of the initial CED^{2}AR system, selected the DDI implementation strategy (based on DDI 2.5), and completed specification for the initial subset of DDI that CED²AR will support. We have also developed a specification for unique dataset references, using Digital Object Identifiers (DOIs), and a strategy to handle the absence of such identifiers in data archive-provided metadata. (b) The goal is to develop boosting and ensemble-based statistical learning techniques to improve the integration, editing and imputation models for various applications in the Census Bureau, for example, assembling the micro-data for longitudinally linked employer-employee database, or predicting multiple responses (e.g., multiple choices such as the race) commonly seen in surveys. We have built the statistical models suitable for multiple responses and we have formulated the solutions using logistic regression as well as tree-based boosting algorithms.

Conferences and papers

- Presented NCRN at the 3rd annual European DDI Users Group Meeting (December 2011), Gothenburg, Sweden
- Presented NCRN at 4th Workshop on Data Access (WDA), March 2012, Luxembourg
- NCRN Poster Presentation at the 38th annual meeting of the International Association for Social Science Information, Service, and Technology (IASSIST) (June 2012), Washington, DC
- NCRN paper accepted at Privacy in Stastical Databases (PSD) 2012
- NCRN paper to be proposed for the next International Digital Curation Conference, (January 2013, Amsterdam)
- Radhendushka Srivastava, Ping Li, and Debasis Sengupta, Testing for Membership to the IFRA and the NBU Classes of Distributions}, AI & Statistics (AISTATS), 2012
- Ping Li, Anshumali Shrivastava, and Christian Konig, GPU-Based Minwise Hashing, International World Wide Web Conference Poster (WWW), 2012
- Sun Xu, Anshumali Shrivastava, and Ping Li, Query Spelling Correction Using Multi-task Learning, International World Wide Web Conference Poster (WWW), 2012
- Anshumali Shrivastava and Ping Li, Fast Near Neighbor Search in High-Dimensional Binary Data: b-Bit Minwise Hashing, Centered and Noncentered Spectral Hashing, and Sign Random Projections, European Conference on Machine Learning (ECML), 2012

Interaction with Census

- Advised Martha Stinson on a process and tools for introducing DDI metadata into the SIPP Synthetic Beta documentation process.
- Discussions with LEHD (Erika McEntarfer) on data documentation and metadata creation for LEHD data.
- Discussed DDI and metadata strategies with Trent Alexander (recently-appointed CES Assistant Center Chief)