Annual Report for Period: 10/2011 - 09/2012
Submitted on: 10/01/2012
Principal Investigator: Abowd, John M.
Award ID: 1131848
Organization: Cornell University
Submitted By:
Vilhuber, Lars - Co-Principal Investigator
Title:
NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation

Project Participants

Senior Personnel

Name: Abowd, John
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Vilhuber, Lars
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Li, Ping
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Block, William
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Lagoze, Carl
Worked for more than 160 Hours: Yes
Contribution to Project:
Support for metadata component.

Name: Brown, Warren
Worked for more than 160 Hours: Yes
Contribution to Project:
Data and metadata support.

Post-doc

Graduate Student

Name: Shrivastav, Anshumali
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Technician, Programmer

Name: Williams, Jeremy
Worked for more than 160 Hours: Yes

Contribution to Project:
Programmer for metadata component

Name: Lee, Camille
Worked for more than 160 Hours: No
Contribution to Project:
Provide support for web interface.

Other Participant

Research Experience for Undergraduates

Organizational Partners

Bureau of the Census
We have collaborated with the Census Bureau, by using metadata provided by the Census Bureau, by providing them with edited metadata, by using the Cornell Research Data Center.

New York Census Research Data Center
We use the Cornell portion of this RDC for our research.

University of Minnesota-Twin Cities
the Minnesota Population Center has provided us with metadata on IPUMS.

Univeristy of Michigan at the Insitute for Social Research
We have collaborated with ICPSR staff (University of Michigan, ISR) on metadata and related issues.

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Findings:

Training and Development:

Outreach Activities:
Presentations of the need for properly curated metadata at field conferences, as a fundamental method of science.

Journal Publications

Radhendushka Srivastava, Ping Li, and Debasis Sengupta, "Testing for Membership to the IFRA and the NBU Classes of Distributions", Journal of Machine Learning Research - Proceedings Track for the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012), p. 1099, vol. 22, (2012). Published,

Ping Li and Anshumali Shrivastava and Arnd Christian K??nig}, "GPU-based minwise hashing: GPU-based minwise hashing", Proceedings of the 21st World Wide Web Conference (WWW 2012) (Companion Volume), p. 565, vol. , (2012). Published, 10.1145/2187980.2188129

Xu Sun and Anshumali Shrivastava and Ping Li, "Query spelling correction using multi-task learning", Proceedings of the 21st World Wide Web Conference (WWW 2012)(Companion Volume), p. 613, vol. , (2012). Published, 10.1145/2187980.2188153

Anshumali Shrivastava and Ping Li, "Fast Near Neighbor Search in High-Dimensional Binary Data", The European Conference on Machine Learning (ECML 2012), p. , vol. , (2012). Accepted,

Xu Sun and Anshumali Shrivastava and Ping Li, "Fast Multi-task Learning for Query Spelling Correction", The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), p. , vol. , (2012). Accepted,

John M. Abowd and Lars Vilhuber and William Block, "A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs", Privacy in Statistical Database, p. , vol. , (2012). Accepted,

Books or Other One-time Publications

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Conference Proceedings

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Activities and Findings: Any Findings
Activities and Findings: Any Training and Development
Any Book

Any Web/Internet Site
Any Product
Contributions: To Any within Discipline
Contributions: To Any Other Disciplines
Contributions: To Any Human Resource Development
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering
Any Conference

Cornell Node Progress Summary September 2012

Comprehensive Extensible Data Documentation and Access Repository (CED²AR)

The NCRN-Cornell node is building a Comprehensive Extensible Data Documentation and Access Repository (CED²AR) designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

Development so far:

- Completed high level functional diagram initial technical diagram of initial CED^{2}AR system (see https://confluence.cornell.edu/display/ncrn/Technical+Documentation)

- Evaluated initial DDI implementation strategy: DDI Lifecycle (3.x) or DDI Codebook (2.5) Settled on DDI Codebook for now; easier to implement and can migrate to future releases as needed. After testing, we decided against using a subset of DDI for the CED^{2}AR: while we may not use all elements, we will store all relevant elements.

- Digitial Object Identifiers (DOI's)

    - Developing formal NCRN Cornell specification for implementing DOI's for datasets.

    - NCRN Cornell will join Datacite, via the California Digital Library. Datacite provides a beta search engine to this metadata repository. The metadata specification for this is an application profile of Dublin Core, and the specification provides a straightforward mapping from DDI 3.12 this metadata format.

- Set up an initial metadata repository, using ICPSR metadata and metadata derived from SIPP Synthetic Beta (SSB) metadata. Subsequent data sets will include QWI, ACS, and IPUMS.

Interaction with Census

- Worked with the SIPP Synthetic Beta team to advise on tools, migrate existing metadata into DDI, and develop tools to maintain the documentation.

Statistical learning and classification

- Modern machine learning techniques for census applications: 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.

- Another useful application is predicting multiple responses (e.g., multiple choices such as the race) commonly seen in survey studies. 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.

Contact

ncrn@cornell.edu