Record Linkage of Web Demographics as a GeoComputation Challenge

T. Edwin Chow [a], Anne H. H. Ngu [b], Yan Lin [a], Clark Phillips [b], Stan Thornhill [b]
[a] Texas Center for Geographic Information Science, Department of Geography
[b] Department of Computer Science
Texas State University – San Marcos

Record Linkage (RL) is the process of matching, resolving and linking multiple records of the same entity from multiple data sources. It is a common technique in database processing and consolidation because duplicated records of the same entity would cause invalid data entries and erroneous inflation. As crowdsourcing Volunteered Geographic Information (VGI) typically engage many observations of geographic features from many users over time and space, RL plays an important role in upholding data integrity, maintaining data quality, and enhancing pattern discovery. The purpose of this positional paper is to highlight research challenges in web demographics and describe relevant research topics.

In the context of web demographics, RL is a GeoComputational challenge because the vast amount of individual observations posts an exponential computational problem to resolve a unique entity. In a data universe with n number of records, each record needs to compare with n-1 other records to examine any duplicated records theoretically. If the VGI include k number of attributes, this RL process would also involve the matching of k attributes independently and their combinations as well.

In general, the first, middle (initial) and last names typically make up a common identifier for a unique individual. However, using only the name for RL is problematic. Common names (e.g. Joe Smith) would require auxiliary information, such as address, phone number, date of birth, associated members, to separate unique individuals. However, such data are often limited in availability and some (e.g. address, phone, etc) can be changed over time. The moving population often produces multiple records with identical attributes but different addresses. The adoption of an English name, a common practice of cultural assimilation among international immigrants in western countries, often results in records with the first and middle names (or initial) that are swapped. Name change among women according to the marital status is a longstanding challenge in surname analysis and demographic studies. It is impossible to get the training data needed for machine learning without field validation of relevant household information. The matching of these cases would require supplementary information and special care in computation to detect, identify, and resolve the entities in questions.

In light of the above challenges, conventional RL approaches use edit distances, such as Levenshtein distance, to count the number of editing operations of string matching in

quantifying similarity. It is interesting to examine the relevance of existing laws and theories to the process of RL in web demographics. For example, Tobler's First Law of Geography (1970) states that everything is related to each other and closer objects are more similar. *Should similar records (e.g. those with identical names but different addresses) that are closer in geographic distance be more likely to be the same person than those that are farther away?* According to Census Bureau (2012), the trend of inverse mobile population with increasing distance holds true among the "local" migration within divisions– 7.72% moved within the same county, 1.96% moved to different counties within the same state, 0.45% moved to different states within the same division, 0.32% moved to different division within the same region. The mobile population increases, however, in regional (0.81%) and international (0.35%) migration. Census data provide useful geographic mobility/migration statistics by races/ethnicity as well as demographic breakdowns. The effectiveness of using such statistics in RL is not known.

According to Zipf's Law (1935), the frequency of an object's property is inversely proportionate to its rank. In urban studies, this implies the population associated with the most common name would be more frequent than other less common names. In the context of web demographics, *should similar records with less common names be more likely to be the same person than more common names?* By applying the same rule to population size, *should similar records in a less populated area be more likely to be the same person than in a populated area?* In addition to the linguistic matching of names, it is interesting to see if the Zipf's Law can be applicable to detect current address among the mobile population with different addresses, *should similar records with an address that are more or less frequent be the current address?*

The research challenges reveal a need to critically evaluate the relevance of existing principles (e.g. Tobler's Law, Zipf's Law) in RL of web demographics. It is essential to carefully implement an appropriate RL algorithm to remove and consolidate duplicate records across multiple data sources and to link web demographics to census data as an independent enumeration alternative. The insights gained from RL of web demographics are relevant to RL of VGI that provides observations, such as Christmas bird counts, extreme weather spotting, etc. More research are needed to understand and capitalize GeoComputation as a way to better process the data, analyze the pattern and interpret the results.

Reference

Chow, T. E., Y. Lin, N.T. Huynh, and J. Davis, 2012. Using Web Demographics to Model Population Change of Vietnamese-Americans in Texas between 2000-2009, *GeoJournal*. 77(1): 119-134.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region, *Economic Geography*. 46: 234–40.

U.S. Census Bureau, 2012. Geographic Mobility/Migration. Geographic Mobility: 2010-2011 – United States < http://www.census.gov/hhes/migration/data/cps/cps2011.html > Accessed June 7, 2012.

Zipf, G. K., 1935. The Psychobiology of Language. Houghton-Mifflin.